

Science and the war on terror

Two years after the attacks on the World Trade Center, the promised reorientation of US national research priorities proceeds without much direction or conviction.

On a beautifully still and sunny September morning two years ago today, the United States received a blow from which the world still reels. It didn't take long for US scientists to realize that 11 September might transform their professional world. The still-unsolved anthrax attacks that followed compounded the impression that the scientific agenda of the nation would be redefined by the challenges of what became known as the 'war on terrorism'.

US researchers recognized that the federal government had a new, overarching mandate. Many set to work to find technical fixes to the formidable array of challenges involved in defending the country from an elusive enemy. The National Academy of Sciences let it be known that its formidable advisory apparatus stood ready to serve the administration of George W. Bush in the nation's hour of need.

Two years on, that expertise has been tapped only sparingly. The government's myriad research agencies have shown varying degrees of success in coming to grips with their roles in the war on terror. The National Institutes of Health has probably turned in the most convincing performance, coping well, in the circumstances, with a huge influx of funds for biodefence research — including \$350 million released last week to build eight regional centres (see page 110).

The redirection of research priorities has already produced results. In biodefence, for example, scientists have made significant progress in identifying and targeting the anthrax toxin, researching into the longevity of the smallpox vaccine, and developing new vaccines, such as the fast-acting Ebola vaccine (see *Nature* **424**, 681; 2003).

The newly formed Department of Homeland Security has visibly struggled to manage its half-a-billion-dollar research budget, however. A *Washington Post* article on 7 September detailed its struggle to attract the high-calibre staff needed to formulate and implement the department's mission, in research or in other spheres. The agency's research wing has so far been able to disseminate only about half of its grant funding this year (see *Nature* **424**, 986; 2003).

Security-related research programmes in other agencies remain diffuse. There has been little sign, for example, that the Department of Energy's huge network of laboratories, which have a strong history

of defence-related work, have been harnessed effectively to meet the new security challenges.

Meanwhile, researchers are finding that they are being affected by the aftermath of the 11 September attacks in unexpected ways — in particular, they are becoming increasingly isolated. Immigration regulations keep scientists from countries such as China, India and Russia away from meetings staged in the United States, so truly international gatherings must take place elsewhere.

Young researchers who want to immigrate to the United States from its closest allies — including France, Germany, Britain and even Canada — are finding it more difficult. The best European universities are enjoying a recruitment mini-boom as a result. It's too early to say how pervasive these developments will prove to be, but their long-term impact on US science could be profound.

On the home front, tighter rules on the handling of pathogens and some examples of clumsy law enforcement — such as the treatment of Thomas Butler of Texas Tech University, who failed to properly explain the destruction of some plague samples from Africa and could face years in jail — may scare researchers away from fields such as biodefence that the government wants to pursue.

And while scientists and the federal government should be marching in step towards the desired goal of better homeland security, there are disturbing signs that the Bush administration remains estranged from the scientific community. Scientific leaders such as Bruce Alberts, president of the National Academy of Sciences, are reduced to writing open letters to public officials complaining about Butler's treatment (see *Nature* **425**, 5; 2003), instead of working quietly to resolve such problems before they arise.

On the global stage, the run-up to the anniversary has been characterized by the Bush administration reaching out to its estranged allies in a belated effort to more effectively pursue a war that the United States cannot win alone. It is high time that the Bush administration also reached out to people inside America — including scientists and their leaders — who feel ignored and increasingly estranged from the direction that the war on terror is taking. ■

Here, there and everywhere

Not before time, there is now a Nature journal devoted to the most diverse branches of the tree of life.

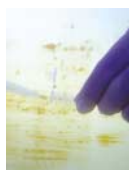
They can survive being entombed for 25 million years in a resin-trapped bee, the conditions inside an operating nuclear reactor, pressures as high as 10 tonnes per square centimetre and depths of 7 km in the Earth's crust. They have existed for most of Earth's history. The number of species is uncountable but is estimated at a billion and more. And although humans depend on the activity of some 30,000 genes encoded in our genome, we are also critically dependent on the presence of up to 4 million genes in upwards of 1,000 species that inhabit the human mouth, intestine, vagina and skin.

Such are the microbes. Nature is pleased to welcome a new sibling that gets to grips with them: *Nature Reviews Microbiology*

(see www.nature.com/reviews/micro). In the multidisciplinary spirit appropriate to a Nature title, the journal seeks to enhance communication between the communities of researchers who study viruses, bacteria, fungi and protozoans.

Its relevance can be expected to extend beyond microbiologists. The journal is collaborating with the World Health Organization and others to ensure strong regular coverage of progress in relatively neglected diseases. And there are topics of broad appeal. How much better might we be at cleaning up polluted environments? What can we infer about the last common ancestor of everything alive on the planet? With such questions addressed in forthcoming issues, a broad readership should take note of the new arrival. ■





Pilot light
Proteomics plan kicks off with preliminary mouse brain study
p110



On the level
Statistics reveal no trend towards summer floods
p111



Future forecasts
Desktop computers unite in bid to tackle climate uncertainty
p112



Youthful trend
Youngsters grow in stature in China's public science arena
p113

Agony for researchers as mix-up forces retraction of ecstasy study

Jonathan Knight, San Francisco

Drug researchers are this week hitting out at scientists who accidentally injected monkeys with methamphetamine — or 'speed' — when they were supposed to be studying the effects of the drug ecstasy.

In results published last September, a team at the Johns Hopkins University School of Medicine in Baltimore, Maryland, showed for the first time that repeated small doses of ecstasy damaged brain cells that produce the neurotransmitter dopamine in monkeys (G. A. Ricaurte *et al.* *Science* **297**, 2260–2263; 2002). Deficiencies in dopamine are linked to neurological disorders including Parkinson's disease. But the researchers will retract their study in this week's issue of *Science*, after concluding that they injected the wrong drug.

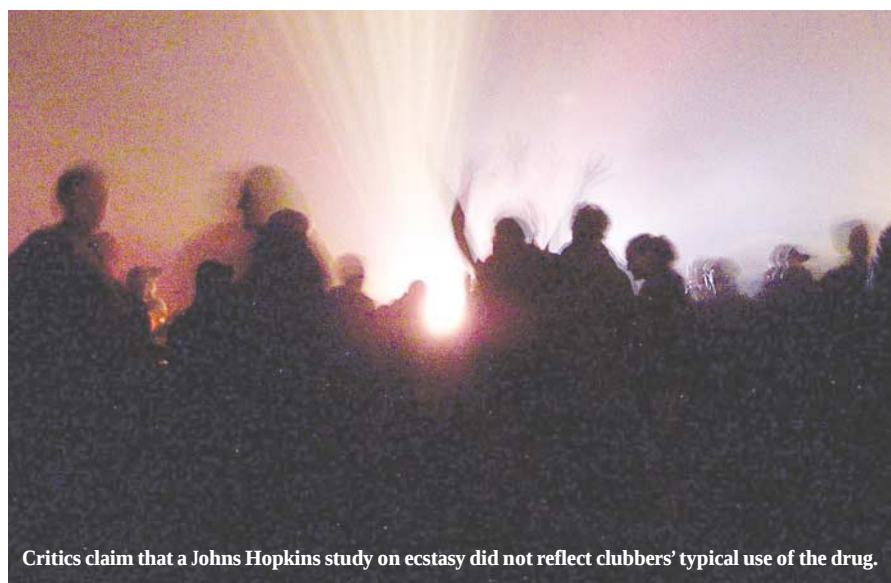
Now critics charge that the team — led by George Ricaurte — was too eager to publish its high-profile finding in this case.

The original study was accompanied by a blaze of publicity in the United States, where it was widely interpreted as evidence that taking ecstasy can lead to brain damage. Johns Hopkins, the National Institute on Drug Abuse (NIDA) in Bethesda, Maryland, which funded the study, and *Science's* publisher, the American Association for the Advancement of Science (AAAS), each heavily promoted the result. Unusually, Alan Leshner, chief executive of the AAAS and former director of NIDA, publicly endorsed the study, saying: "It sends an important public-health message — don't experiment with your own brain."

In their retraction, the authors say that they have been unable to repeat their results. After months of trying to trace the problem, they concluded that all but one of the monkeys had actually received methamphetamine, a powerful stimulant known to damage dopamine-making cells.

The mix-up occurred before the drugs arrived at the lab, the authors say. Similar quantities of the two drugs were ordered and arrived on the same day, their labels apparently reversed. Although the original vial labelled ecstasy has been discarded, tests on the methamphetamine vial show that it contained pure ecstasy, the retraction says.

But the authors should have been suspi-



Critics claim that a Johns Hopkins study on ecstasy did not reflect clubbers' typical use of the drug.

cious sooner, says Charles Grob, a psychiatrist at the Harbor-UCLA Medical Center in Los Angeles who studies ecstasy, also known as MDMA. "It was implausible that MDMA could have caused those findings," he says. Critics have argued that the study, in which the injections resulted in the sudden death of two out of ten monkeys, could not be representative of recreational ecstasy use by people.

Grob and other critics have in the past argued that other studies by Ricaurte, drawing on brain scans, were equally inconclusive. "This is part of a pattern in Ricaurte's work," Grob alleges.

John Henry, a psychoactive-drug researcher at Imperial College, London, characterizes the retraction as a carefully worded plea for more money. He points out that the last sentence leaves open the possibility that the basic result is right, despite failed attempts to repeat the experiment. "It's extremely arrogant," he says. Henry claims that NIDA favours studies that are liable to prove the toxicity of recreational drugs.

Ricaurte could not be reached for comment but his wife and co-author, Hopkins researcher Una McCann, defended the basic hypothesis that ecstasy use can cause Parkinson-like symptoms. She says that it is

clearly toxic to dopamine cells in mice, for instance. "We thought we were getting the high levels of dopamine toxicity seen in other species," she says.

Michael Taffe, who studies MDMA at the Scripps Research Institute in La Jolla, California, says that it would have been a tough call whether to publish the original paper, given the potential importance of the finding. Nor would anyone normally test the contents of a labelled bottle from a manufacturer: "I don't know what I would have done," he says.

Ricaurte's group ordered the drugs from NIDA, which supplies approved researchers through a contractor, RTI International of Research Triangle Park, North Carolina. RTI said in a statement that it was reviewing procedures, but has no evidence that it had made a mistake.

NIDA associate director Timothy Condon says that NIDA will assist RTI in the review, but that there would be no further investigation of how the mix-up occurred. "We have asked RTI to look at their procedures," he says. "I don't know what else we could do."

But Grob doubts whether there will be a real attempt to get to the bottom of the episode. "I think NIDA needs to answer some questions," he says.

Brain protein project enlists mice in 'dry run'

Alison Abbott, Munich

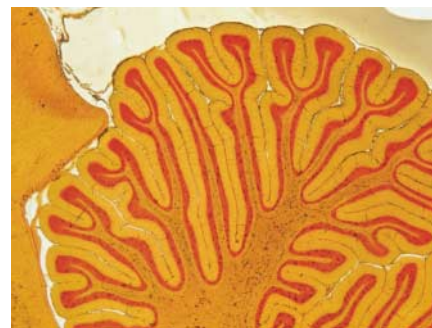
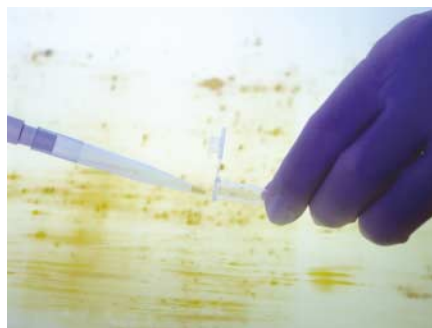
A major global project to catalogue the proteins in the human brain has been launched with a pilot that will analyse proteins expressed in the mouse brain.

Scientists from around the world agreed on the pilot phase at a meeting in Düsseldorf, Germany, on 5–6 September, and decided that the long-term aims of the project should focus on proteins that are relevant to Alzheimer's and Parkinson's diseases.

But they also discussed the technical and scientific problems facing the Human Brain Proteome Project (HBPP) as it aims for its final goal of cataloguing the protein-expression profile, or proteome, of the human brain in normal health, in disease and in old age.

Forty-five scientists from ten countries, mostly in Europe, attended this first HBPP meeting. The project is being led by protein chemists Helmut Meyer from the University of Bochum and Joachim Klose from Humboldt University in Berlin, who are co-directing the world's largest brain-proteomics project to date, begun in 2000 with €15 million (US\$17 million) in funding from the German ministry of research.

The HBPP concept was launched in April under the umbrella of the Human Proteome Organisation (HUPO), the global group for public proteomics initiatives. It joins existing HUPO initiatives to determine the



Hands-on heads: proteomics researchers hope to sample protein levels in the human brain (right).

proteomes of human plasma and liver.

All of these initiatives must cope with the difficulties of measuring thousands, or even hundreds of thousands, of proteins, some of the most biologically important of which may be present at very low levels and buried in lipid membranes. But the brain project also has the unique problem of requiring access to fresh, healthy human material. To begin the project, the HBPP needs 10–20 human brains.

Autopsied brains are a potential source of tissue, and have the advantage of being 'normal' if they come, for example, from otherwise healthy road-accident victims. But autopsy material is never fresh, and many proteins will already have been degraded if an autopsy is performed more than a few hours after death.

"The problem of degradation of brain

proteins after death has dogged brain researchers for decades," says Hans Kretzschmar, a neurobiologist at the University of Munich and the coordinator of BrainNet Europe, a network of ten brain banks. "No one yet knows just how bad it is — and just how much it will affect our project."

The HBPP will investigate the rates of protein degradation after death, says Meyer. In the meantime, Kretzschmar says he and his colleagues will try to acquire post-mortem brains no later than three hours after death, the earliest that coroners can proceed with an autopsy after receiving permission from relatives, which is required in western Europe.

The one-year mouse pilot project will analyse proteins in the whole brains of a commonly used lab strain; 16-day-old embryos, 7-day-old mice and 8-week-old adult animals will be investigated. The project will catalogue the proteins and compare different methods used by the various HBPP labs — of which there are around 50 so far, mostly in Europe — for their ability to reliably quantify differences in the expression of 5,000–10,000 proteins in each age group. "We'll be looking at all the proteins we can with current technologies — and this includes some of the low-abundance ones, not just the easy ones," says Meyer.

Work on healthy human brains will begin as they become available. Proteomes will be analysed in four relevant brain areas rather than in the whole brain. Two of the areas — the temporal lobe and the hippocampus, which can also sometimes be acquired from surgery — are important in Alzheimer's. The other two — the substantia nigra and the basal ganglia, which cannot be obtained from surgery — are important in Parkinson's. These will eventually be compared with autopsied material from Parkinson's or Alzheimer's patients. But given the limitations on obtaining human material, much of the work will first be done in mouse models of the diseases.

Quality-controlled data from the HBPP will be made publicly available on the Internet, in accordance with HUPO's philosophy. The project's participants will seek funds from local and international sources. ■

■ www.hbpp.org

US selects regional biodefence hubs

Rex Dalton, San Diego

The US National Institutes of Health (NIH) has awarded \$350 million in grants to eight university consortia to set up research centres that will spearhead the country's biodefence research effort.

The consortia will each receive between \$35 million and \$50 million over five years from the NIH's National Institute of Allergy and Infectious Diseases (NIAID) to boost research programmes on defences against biological weapons.

Anthony Fauci, the NIAID's director, says: "The new programme provides a coordinated and comprehensive mechanism to support the interdisciplinary research that will lead to new and improved therapies, vaccines, diagnostics and other tools to protect the citizens of our country and the world."

By the end of this month the NIAID will also announce one or two institutions that could receive as much as \$120 million to build biological containment facilities for research on the deadliest pathogens.

In the aftermath of the terrorist attacks

of 11 September 2001 and the subsequent anthrax mailings, the NIAID boosted its bioweapons defence programmes by increasing grants and inviting institutions to host the new research centres.

About 15 consortia applied for the scheme. The winners, which were announced on 4 September, are led by Duke University in Durham, North Carolina; Harvard Medical School in Boston, Massachusetts; the New York State Department of Health; the University of Chicago and Northwestern University, both in Illinois; the University of Maryland in Baltimore; the University of Texas Medical Branch at Galveston; the University of Washington in Seattle; and Washington University in St Louis, Missouri.

"While this project has been driven by concerns about bioterrorism, the knowledge gained could have a significant impact on humanity's eternal battle against all infectious diseases," says geneticist Olaf Schneewind of the University of Chicago, one of the programme's principal investigators.

Analysis pours cold water on flood theory



Trend free? This year's low water levels are in stark contrast to last summer's floods on the Elbe.

Quirin Schiermeier, Munich

Last summer's flooding around the River Elbe and its tributaries was one of central Europe's biggest natural disasters in living memory, and some observers saw it as part of a trend towards more extreme weather. But according to a comprehensive analysis by German climate researchers, there is no evidence that this is so (see page 166).

The researchers' statistical analysis of recent river-flow data and historical records shows no upward trend in the frequency of extreme summertime flooding in eastern Germany, Poland and the Czech Republic.

Records tapped by the analysis include the Weikinn archive, a bank of more than 20,000 documents on floods and rainfall dating back to the eleventh century. The archive — compiled by the late Curt Weikinn, a self-taught climatologist from the former East Germany — has often been ignored because

of the archivist's unsystematic approach. But, the fresh analysis finds, a comparison with other climate records suggests that the archive's flood data are very reliable.

The analysis finds that changes in temperature over the past 1,000 years have affected hydrology in the region, but winter flooding actually decreased over the past century, and summer flooding shows no trend.

"According to the records, the Elbe flood last year was indeed a centennial event, with an average return period of 168 years," says Manfred Mudelsee, a hydrologist at the University of Leipzig, and the paper's lead author. "We had almost finished our analysis when the flood came, and we decided to include it. But it would be frivolous to derive from this single event any kind of statistical trend for the past 50, 100 or 150 years," he says.

Analysis of historic sources is worthwhile and instructive in climate research, provided

that the findings are interpreted with due caution, says Axel Bronstert, a hydrologist and flood expert at the University of Potsdam in Germany. But he warns against interpreting the findings as an indication that flood occurrence won't increase in the future.

"The absence of an upward trend might appear reassuring," he says, "but it isn't." At least one high-resolution climate model suggests that, on the contrary, summertime flooding in Europe may become more frequent under the influence of greenhouse-gas-induced global warming, even though continental summers are becoming drier (J. H. Christensen and O. B. Christensen *Nature* **421**, 805–806; 2003). The latest findings don't disprove these predictions, Bronstert says.

According to the insurance company Munich Re, the Elbe flood caused US\$18.5 billion in damage. In the wake of the flood, Germany and the Czech Republic agreed on a cross-border action plan to improve flood warning and protection systems.

The plan, which includes the creation of larger flood plains to buffer peak flows, remains necessary, says Christian Korndörfer, head of the environmental office of Dresden, the city worst hit by the flood. "Whether or not you can make out a trend, the Elbe last year showed us what is possible, and we will have to take appropriate precautions," he says.

Flood managers take into account how climate change may affect precipitation and runoff, says Otto Malek, a flood-protection expert at the German federal environment ministry. But given the risks and prevailing uncertainties, it would be negligent to delay concrete actions until we have more robust trends, he adds.

"The real problem is that there are too many people building houses, and amassing material assets, in areas known to be in danger of flooding," Malek says. ■

Ministers appointed to engineer revival in Iraqi research

Geoff Brumfiel, Washington

The rebuilding of Iraq's scientific infrastructure stepped up a gear last week with the establishment of a research ministry.

Iraq's chief civilian administrator, US diplomat Paul Bremer, announced on 1 September that the US-led coalition occupying Iraq had abolished the Ministry of Atomic Energy and set up a Ministry of Science and Technology in its place. The new ministry features six directorates dealing with energy, the environment, information technology, agriculture, materials science and industrial development. It will be led by

Rashad Omar Mandan, a British-trained civil engineer who served at the Iraqi oil ministry until he fled the country in 1999.

Mandan was sworn in as minister of science and technology on 2 September along with Iraq's first minister of higher education, Ziad Abderrazzak Mohammad Aswad, a US-trained engineer who previously headed the oil-engineering department at Baghdad University.

The US government has also named John Agresto, former president of St John's College in Sante Fe, New Mexico, as the senior American official who will oversee Iraq's

university system. "From what I gather, the situation is very bad — but not desperate," Agresto says. He adds that his priorities will be to rebuild the infrastructure of the colleges and hire faculty members who were dismissed or exiled under former Iraqi president Saddam Hussein.

Agresto, a political scientist, was deputy head of the National Endowment for the Humanities under President Ronald Reagan. During his tenure at St John's College, he got to know Donald Rumsfeld, the US defence secretary, whose wife sat on the college's board of trustees. ■

US reflects on flying eye for transgenic crops

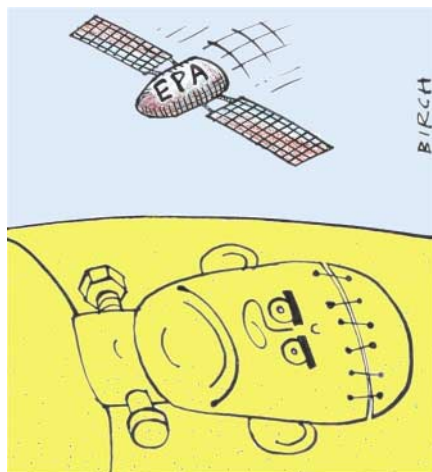
Jonathan Knight, San Francisco

The US Environmental Protection Agency (EPA) is hatching a scheme that could let it monitor genetically modified crops from space. Experiments will begin next spring to determine whether subtle differences in the way leaves reflect the Sun's rays can distinguish transgenic from conventional maize.

If it works, the technology would allow the EPA to track farmers' compliance with planting guidelines, and might even spot the emergence of insecticide-resistant pests. This would be a remarkable feat of detection because, to the naked eye, transgenic plants appear identical to normal ones.

Satellite imagery is already a well-established agricultural tool. Stressed plants absorb less infrared light than healthy plants, so they appear 'warmer' in infrared photographs. Farmers already make use of images from satellites and aeroplanes to spot thirsty crops, pest infestations and plant diseases. Some large-scale growers use infrared data to classify weeds in their fields to help them decide which herbicides to apply.

All of these things can also be spotted by a farmer in the field — the main advantage of observations from the air is that they allow large areas to be examined quickly. But the EPA's scheme to monitor transgenic crops by satellite would involve discerning more



information about a plant from space than an observer could do on the ground.

The idea rests on the possibility that subtle genetic differences between plants might influence the spectral qualities of the solar radiation reflected by the leaves, says project leader John Glaser of the EPA's Cincinnati office. Transgenic varieties would appear different, not because of their inserted genes, but because the starting strains from which they were generated were different from those of conventional crops, he says.

To achieve this sensitivity, computer

models will be needed that can peel away background noise, such as the effects of rainfall, air temperature and pest attack. With this in mind, Glaser has recruited agricultural experts at Pennsylvania State University and the University of Nebraska in Lincoln.

The researchers admit that this may be easier said than done. "The big question is, can we refine it enough to see genetic differences?" says project collaborator Joe Russo, president of ZedX, a company in Bellefonte, Pennsylvania, that builds computer models for agriculture.

The project was born mainly out of concern that overuse of *Bt* maize (corn) — which is genetically modified to produce a natural insecticide — will result in the development of resistant insects, rendering the technology useless. US farmers who plant *Bt* maize are required to keep conventional maize on 20% of their acreage to minimize the risk, but up to one-fifth are thought to be flouting this rule, according to a study released in June (see *Nature* 424, 116; 2003).

In theory, the project could also spot fields of *Bt* maize that are under attack by insects, helping regulators to nip resistance in the bud. The technology, which would not necessarily be limited to maize, could ultimately help to track the use of transgenic crops worldwide, Glaser says. ■

Global effort to plot climate change

Heike Langenberg, London

A major source of uncertainty in models of climate change could be reduced by a project utilizing the spare power of desktop computers.

The backers of climateprediction.net, due to be launched on 12 September, are inviting the public to download software that will help to probe the assumptions used in sophisticated models of Earth's atmosphere. By running thousands of these models simultaneously on home and office machines, researchers hope to identify the ranges of assumptions that are realistic, and hence to generate better predictions of future climate.

Climate models use rules to describe how large-scale factors such as temperature and humidity determine smaller-scale properties such as cloud cover. A range of rules are compatible with observations, so researchers pick the ones that most accurately make the models follow the course of climate in recent decades.

According to Myles Allen, a climate researcher at the Rutherford Appleton Laboratory near Oxford and the originator

of climateprediction.net, it is important to explore the full range of possible rules in order to understand the range of realistic future-climate scenarios.

With time on climate supercomputers scarce, this has not been done. But other big tasks, such as the analysis of extraterrestrial radio signals for signals from intelligent life forms, are already being performed by parcelling out the task to large numbers of personal computers. "Distributed computing projects have attracted hundreds of thousands of participants," says Allen. "Such an ensemble size would form an extraordinary experiment."

Slightly simplified versions of a model that has been developed at the Hadley Centre for Climate Prediction and Research in Bracknell, UK, will be available for downloading from the project website. Each model uses a unique set of rules and simulates a 15-year period, during which the stability of the rules is checked. It then simulates 15 years of current climate conditions, and finally a world with increased greenhouse-gas concentrations. This should produce better estimates for a



Future in their hands: children at Gosford Hill school in Oxfordshire examine a simulation.

range of parameters such as temperature increase, and allow scientists to single out the rules for determining the small-scale processes to which climate is most sensitive.

Allen aims to generate publicity at this week's launch, which takes place at the Science Museum in London and coincides with the British Association's annual science festival. The first phase of data collection should be completed by the end of the year, he says, and the first results could be available early in 2004. ■

♦ www.climateprediction.net

Young guns shoot to the top in China's research revolution

David Cyranoski, Beijing

As older scientists retire and research rapidly expands, a large number of young scientists in China are being thrust into prominent positions.

This year, for example, the Institute of Biophysics in Beijing, part of the Chinese Academy of Sciences (CAS), has hired 13 new group leaders with an average age of less than 40. The institute's director, 52-year-old Zihao Rao, says that he plans to bring in another 20 young group leaders as most of the current senior investigators retire over the next couple of years.

Other CAS institutes and some university departments are adopting a similar strategy. China's readiness to thrust researchers into senior positions in their 30s — when, it is often said, most great scientists do their best work — contrasts with the trouble that other

Information Centre, presented data at the meeting which show that as recently as 1995, there was a 'hump' of older scientists whose training predated the turmoil of the late 1960s. Now that older generation has retired, leaving a demographic that is dominated by youth. "This is clearly a result of the Cultural Revolution," says Jin.

Now, as the post-revolution scientists are coming into their prime, the Chinese government has taken a keen interest in promoting science, encouraging institutional leaders to invest in talented young researchers. "They gave me the green light," says Rao, many of whose recruits have only recently returned to China after travelling abroad for training. The United States, for example, welcomed students who left China in the wake of the failed revolt against the government in Tiananmen Square in 1989.

Enge Wang, director of the CAS Institute of Physics, says that his group leaders — two-thirds of whom are still in their 30s — are bursting with youthful enthusiasm. "Half of them will be at the laboratory until 11 at night. It's a very open and energetic atmosphere," he says.

Much of the talent that is emerging now resulted from China's struggle to rebuild its intellectual élite during the 1980s and 1990s. Government projects, with names such as the National Distinguished Youth's Science Foundation and the One Hundred Talent Project, were set up to cultivate young scientific talent, and to attract trained researchers back home from the United States and Europe, Jin says. As these projects bear fruit, the country's premier research institutions are raising their standards. In the past six years, for example, the CAS and Tsinghua University have begun to insist that newly hired professors have PhDs.

Zhu Chen, a biologist and vice-president of the CAS, says that he is happy to see all this youthful vigour, but is also a little apprehensive about giving promising people, such as Rao's 33-year-old deputy director, too much responsibility. "It would be better to have some accomplished leaders in their 50s and 60s to let the young people do their work," he argues.

Jin's data also forecast another cloud on the horizon. As of the late 1990s, a dip in the number of researchers in their late 20s and early 30s suggests that a large number of young scientists have moved overseas. But if the past is any guide, many of these researchers will still play a role in China's flourishing research community, either by returning to their homeland or by collaborating with scientists who work there. ■



Youth culture: students at Peking University help to feed the expansion of China's science base.

scientific powers, including Japan and the United States, have experienced in creating opportunities for young researchers in their prime (see *Nature* 421, 680; 2003).

The fast track being taken by many young researchers in China is a product of two factors: the rapid expansion of research in the country, and the lack of established, middle-aged scientists from a generation whose best talent was squandered during Chairman Mao's Cultural Revolution, which began in 1966. There followed ten years during which scientists and other intellectuals were sent to labour on collective farms.

Surveys of the age distribution of Chinese researchers, presented two weeks ago at the ninth biannual International Conference on Scientometrics and Informetrics in Beijing, clearly demonstrate the revolution's impact on China's scientifically active population. Bihui Jin, of the CAS Documentation and



Pattern pending: Stephen Wolfram takes his 'new kind of science' to Washington.

Esoteric theorist lands starring role in Senate hearing

Geoff Brumfiel, Washington

For the past 15 months, mathematician Stephen Wolfram has been packing university lecture halls with people eager to hear about his 'new kind of science'. But on 4 September, he addressed an audience unlike any he's had before — a US Senate subcommittee, whose chairman has expressed an interest in providing public funding for his work.

Wolfram made waves in May 2002, when he published *A New Kind of Science*. The book describes how feeding simple sets of rules into a computer can generate remarkably complex patterns. The author theorizes that these results can be used to describe virtually everything from the patterns on mollusc shells to the fabric of space-time (see *Nature* 417, 216–218; 2002). Scientists, however, have given the book mixed reviews.

To date, Wolfram has privately funded his research and the publication of his book. But the publicity surrounding *A New Kind of Science* has caught the attention of Sam Brownback (Republican, Kansas), who chairs the Senate commerce committee's subcommittee on science, technology and space.

Brownback is one of the more conservative members of the Senate, and has clashed with scientists on issues such as embryonic-stem-cell research and the teaching of creationism in the classroom. He was the only senator to attend the hour-long hearing, at which Wolfram explained why he believes that his computer-generated patterns might benefit the fields of physics, biology and nanotechnology.

It is unclear which US research agency might be responsible for following this up, but Brownback seems committed to finding a home for Wolfram's research. "It seems like this work could have enormous potential," he told *Nature* after the hearing. "If it's right, then the government should be investing in it." ■

India gives go-ahead for fast-breeder reactor project

New Delhi The Indian government is to spend 35 billion rupees (US\$760 million) on a fast-breeder nuclear reactor — a form of power generation that other countries are abandoning.

India's nuclear reactors burn natural uranium, but its reserves will run out in 25 years. It has lots of thorium, however, which the government says it wants to convert into a fissile isotope of uranium using fast-breeder reactors. The reactors would burn natural uranium and spent nuclear fuel, producing the energy and fast neutrons needed to transform thorium into uranium.

Some nuclear-power experts argue that fast-breeder reactors are uneconomical, and warn that plutonium is far more toxic than uranium. But Indian officials say they have a decade of experience with a test breeder reactor, although no other country has had much success. France shut down its problem-ridden Superphénix reactor in 1997, and Japan has not restarted its Monju reactor since it was closed down after a fire in 1995.

The plant — the first of four planned fast-breeder reactors — will be built in Chennai and is expected to be complete in 2011.



Sail away: solar sails could provide the propulsion needed to reach interstellar space.

Compass test helps solar sailors reach for the stars

Washington A US company's bid to send a commercial solar sail into interstellar space received a boost from NASA last week, when the agency selected the firm's craft to test a new navigational instrument.

Team Encounter of Houston, Texas, aims to send a solar sail beyond the Solar System. The craft — propelled by a stream of photons from the Sun — will carry photographs, messages and DNA from individuals who pledged funds. A test flight, scheduled for launch into Earth orbit in 2005, will carry NASA's Inertial Stellar Compass, which uses a gyroscope and images of stars to provide

navigational information. NASA will pay Team Encounter \$6.5 million for the launch.

This is not Team Encounter's first government contract — the US National Oceanic and Atmospheric Administration previously asked the company to help to design solar-sail craft to fly in Earth orbit.

♦ www.teamencounter.com

Antarctic data reveal hole truth about ozone

Paris The ozone hole over Antarctica could soon be larger than ever. The Geneva-based World Meteorological Organization announced last week that the hole is now 25 million square kilometres, and is on track to break the record of 28.3 million square kilometres seen in 2000.

The organization attributes the ozone loss to weather conditions similar to those in 2000, when the first sunlight of the Antarctic spring accelerated ozone-destroying reactions in the stratosphere while temperatures remained low. Ozone levels should recover as temperatures rise from October onwards.

The size, depth and persistence of the ozone hole vary from year to year — last year's hole was the smallest since 1988 — and a single year cannot be used to predict general trends. Atmospheric-chemistry models predict that ozone levels over the

continent will hit a low some time before 2010, recovering to 1980 levels by the middle of the century thanks to the phasing out of ozone-destroying chemicals.

US cuts use of lie detectors on job-hunting researchers

Washington The US Department of Energy is reducing the number of researchers who will be subjected to polygraph examinations. The move comes in response to a 2002 study by the National Academy of Sciences, which concluded that "polygraph testing yields an unacceptable choice ... between too many loyal employees falsely judged deceptive and too many security threats left undetected".

The energy department, which uses polygraph tests to screen researchers working on sensitive research, initially said it would continue to use the lie detector in job interviews (see *Nature* 422, 793; 2003). But in a Senate hearing on 4 September, deputy energy secretary Kyle McSillarow announced a reduction in the number of positions that require screening from 20,000 to 4,500.

Researchers opposed to screenings maintain that this number is still too high, and say they will continue to fight to have polygraph screening removed completely from the department's laboratories.

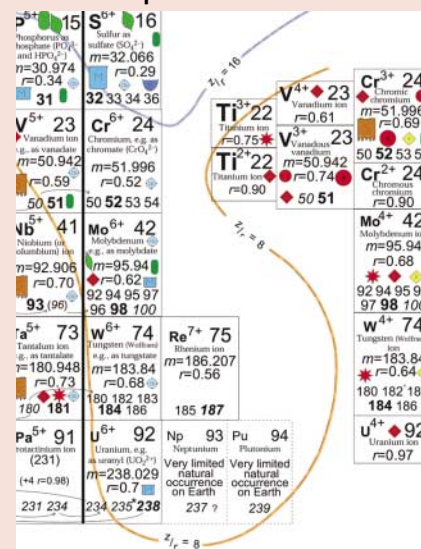
Ion charges to heart of Earth's periodic table

London Geologists now have their own periodic table: a fresh layout that better reflects how different substances are distributed in nature.

Bruce Railsback, a geologist at the University of Georgia in Athens, Georgia, redesigned the table using ions rather than atoms. The result, a small sample of which can be seen here, is a table in which ions group together according to their chemistry and whereabouts on Earth. Families are based around groups such as trace minerals in soil and chemicals dissolved in surface water. The patterns more or less follow the interactions of each ion with oxygen.

The table doesn't explain the science behind the trends, but geologists have welcomed the colourful *aide memoire*. "It's a work of art," says Stephen Elphick, a geologist and geophysicist at the University of Edinburgh, UK.

♦ www.gly.uga.edu/railsback/PT.html



Second astronaut given NASA's top science job

Washington NASA has appointed John Grunsfeld, a veteran of four space flights with a background in high-energy astrophysics, as its chief scientist. It is the second successive time that the agency has selected an astronaut for the position.

Grunsfeld twice serviced the Hubble Space Telescope in orbit, and recently advocated extending Hubble's life (see *Nature* 424, 603; 2003). He has used the Compton Gamma Ray Observatory and the Very Large Array in New Mexico for his research.

Grunsfeld will take over from Shannon Lucid, who is best known for her six-month stay on the Mir space station in 1996.

Biotechnology at the bar

Science is moving too fast for the legal system to keep up. But lawyers and scientists have a solution — a body that would help courts tackle cases involving the latest research. Nicola Nosengo investigates.



Italian judge Amedeo Santosuosso and biologist Carlo Alberto Redi (far right) talk to judges and scientists at the University of Pavia.

This October, Franklin Zweig will face one of the most difficult cases in his career as a judge. A radioactive, mercury-laden waste sludge has been deposited on both banks of a river that forms the border between two countries. The heavy metal is seeping into nearby water sources, which supply hundreds of thousands of people.

One country wants to apply a genetically modified, radiation-resistant bacterium that can immobilize the mercury. But on the other bank, activists who fear the transgenic bacterium more than radioactivity and heavy-metal poisoning have asked a court to block the project. "How can a judge obtain the scientific information that is necessary to resolve such a case?" asks Zweig.

Luckily, this case is hypothetical — an imaginary scenario designed to test the International Science and Technology Reference Forum, an ambitious initiative developed in response to the increasingly difficult relationship between the law and science, particularly fast-moving fields such as biotechnology. "Discussions between scientists and legal experts have reached a common conclusion," says

Zweig, a former legal adviser to the US Senate. "A new institution is necessary to mediate between science and the legal environment."

Hard cases

Advocates of the forum hope that by providing a permanent team of legal experts and scientists, it will help national courts to resolve disputes like the one described above. Independent advice is needed, says Zweig, because biotechnology is forcing judges to deal with science in a way that they never had to before. He is currently president of the Einstein Institute for Science, Health & the Courts, a non-profit organization in Bethesda, Maryland, which is promoting the new initiative.

Take cloning, for example. If a human is ever cloned, courts could potentially face tortuous questions about the legal relationship between an egg donor, her husband, the person being cloned, a surrogate mother and the resulting child. They might also have to decide what legal rights, privileges and immunities a cloned child could claim in a jurisdiction that bans human reproductive cloning.

Such issues should be resolved in advance by national parliaments, says Zweig. But parliaments are generally slow to act, so policy decisions about emerging technologies are often made in court. This was the case with the long-debated issue of whether patents could be granted for living organisms obtained through artificial genetic modification.

In 1972, Ananda Chakrabarty, a microbiologist at the University of Illinois at Chicago, applied for a patent on a genetically modified bacterium that could break down crude oil (A. M. Chakrabarty *et al. Proc. Natl Acad. Sci. USA* **70**, 1137–1140; 1973). The US Patent Office rejected his request, saying that living things could not be patented. An appeal court reversed the decision, and the case reached the Supreme Court in 1980. In a landmark decision, the court awarded Chakrabarty the patent, declaring that "everything under the sun that is made by man is eligible for patenting". The decision, which was influential in shaping US biotech policy, was made before any specific law had been drafted to address the issue.

The Einstein institute was born just over a



Trading places: Judge Valentina Sellaroli loads a DNA gel (top); Ananda Chakrabarty with the first patent for a living organism, issued in 1981.

decade later, after even more responsibility for scientific law passed into the hands of US judges. In 1993, the Supreme Court imposed upon judges the duty of assessing the validity of novel scientific evidence. One of the new institute's first activities was to provide education programmes on genetics for judges. Since then, the institute has run scientific courses for more than 3,000 judges in the United States, Europe and Asia.

Institute staff soon realized that science was forcing the legal system to depart from its usual mechanisms. Scientific cases often force judges to look forwards rather than back, for example. "Judges typically study previous cases," Zweig explains. "To cope with science, they must try to forecast the future." Science also forces courts to think internationally, as certain issues, such as the spread of genetically modified crops, can be difficult to consider on a national level. "We faced a paradox," says Zweig. "Law is by definition local, but science is by definition global, and so are the legal problems that new technologies inspire."

An informal group of scientists and judges evolved to discuss such issues, and the forum has been developed in part to transform this community into a new international judicial institution. Its members hope that it will act as a kind of supreme court for complex cases involving scientific matters that national courts fail to resolve. If a national court decides to refer a case to the forum, one group of forum members will discuss the case, and their verdict will be reviewed by a separate team from within the forum. Their decisions will be founded upon analyses of the science and technology involved, risk assessment and the ethical and religious values that shape national legislations.

In addition to the courts, private parties and administrative, regulatory and legislative bodies will be eligible for the forum's assistance. Client courts will be free to use or ignore the forum's advisory verdict in resolving a high-profile dispute, and the forum itself will not seek to enforce its decisions. "Its legitimacy flows from independent, neutral knowledge power, not a police power," Zweig says.

The idea is now moving from concept to reality. The United Nations has already agreed to fund the initiative. Next month, the panel from the Einstein institute that is developing the forum will meet at the Missouri Botanical Garden in St Louis. With its collection of transgenic plants, the garden is a perfect setting for the first steps of an institution covering biotechnology law. Simulated case scenarios will be assessed, and foundation documents critiqued. The forum's provisional governing council is due to be elected in March 2004, and the forum could be operational a year later. It is likely to comprise a permanent team of judges, with scientist members acting as part-time consultants.

Independent view

According to Chakrabarty, who is chief scientific adviser to the project, scientists have just as many reasons to be interested in the forum as judges have. The organization might, for example, be asked by governments to produce a neutral assessment of the benefits and costs of technologies such as embryonic stem cells.

But not everyone is convinced that it is the best way forward. Italian judge Amedeo Santosuosso sits on the organizing panel and says that the forum could be particularly useful outside Europe and North America, where courts do not always have access to good scientific experts. But he fears that as a voluntary, non-binding institution, it could face difficulty in getting its verdicts widely accepted.

The rules on its composition, which are still to be decided, will be crucial, he says. He thinks that the forum will need a broad range of participants if it is to be perceived as neutral, independent and fully representative, but fears that too many members will make it inefficient. Potential problems like these are to be discussed at the St Louis meeting.

In the meantime, Santosuosso is promoting the European Network for Life Sciences, Health and the Courts, formed this June, and based at present at the University of Pavia in Italy. The network brings together legal experts and scientists from European countries, and organizes courses, seminars and conferences where judges and scientists can meet. Together with Carlo Alberto Redi, a developmental biologist at the University of Pavia, Santosuosso recently organized a science course where 12 judges spent a week at Redi's laboratory, working side-by-side with researchers and becoming acquainted with techniques such as DNA testing.

Bench law

"We were trying to make judges aware of what a scientist's work actually is," says Redi, "so that when they have to decide whether an expert opinion is acceptable, or whether an experiment is breaking the law, they know what goes on in a laboratory." Redi speaks from experience. In 1998, after members of his laboratory participated in the experiment that led to the first cloned mouse (T. Wakayama *et al.* *Nature* **394**, 369–374; 1998), he found the police waiting at the lab. At the time, Italian legislation banned any form of cloning, and the fact that the experiment took place in Hawaii did not spare Redi's team from police questioning.

The initiative will be extended elsewhere in Italy, and similar courses will start next year in Germany, Spain, Norway and Switzerland. "In the long run, smaller initiatives can be even more effective than the international forum," Santosuosso says. "They create opportunities where the world of life sciences and the world of law actually work together."

While acknowledging the importance of local initiatives, Chakrabarty points out that problems that defy national boundaries are best addressed by a permanent international body. He also acknowledges another potential problem: some parties may refuse to use the forum if they think scientific opinion will not back their case. Chakrabarty says this likelihood should be minimized if the forum tries to acknowledge all shades of scientific opinion, and allows the scientific evaluations to be filtered by its judicial members, who will take into account the economic, ethical and political angles of a dispute.

So will the forum succeed? With United Nations support, and a track record of dialogue between the legal and scientific communities, its backers are well placed to make sure it does. Zweig, for one, is in no doubt about the need for the forum. "Biotechnology is pushing society into a new area," he says, "where the letter of the law is grey, not black."

Nicola Nosengo is an intern in Nature's Munich office.

Einstein Institute for Science, Health & the Courts

◆ einshac.org/EINSHAC_FRAME.html

European Network for Life Sciences, Health and Courts

◆ www.unip.it/Biol/ENLSC.html



The ups and downs of lithium

This pill contains lithium. It's been used to treat manic depression for decades, and may help combat other brain disorders. So how come no one knows for sure why it works? Helen R. Pilcher reports.

"He sleeps normally, eats normally," says Sarah. "He's a fairly typical kid." Her son is one of millions of patients with bipolar disorder (BPD) — manic depression — whose lives have been transformed by lithium. Children with the condition can see-saw between energetic highs and listless lows many times a day. Adults suffer similar swings, albeit at less frequent intervals.

During periods of mania, sufferers can experience euphoria and exercise poor judgement, perhaps going on spending sprees. Depressive episodes bring on fatigue, pessimism and anxiety. But lithium, administered as a salt, puts sufferers on an even keel, freeing some parents from the need to provide constant care for children with BPD. "One minute he'd be bouncing off the walls and giddy, the next minute he'd be crying," recalls Sarah. "Now, he has settled down with himself and can play with other children."

Despite being the most widely used treatment for BPD for over 30 years, lithium is an enigma: no one knows how it works. Even though it can transform lives, little funding is being channelled into clinical studies of the drug. Lithium cannot be patented, so drug companies are investigating more profitable alternatives. Many doctors are

increasingly avoiding the treatment, fearing that it is difficult to prescribe. Lithium, it seems, is in need of some good PR.

Help may be at hand. Recent results have linked the drug to an enzyme involved in a range of cellular processes. The new studies could explain why lithium works, and provide hints about how it could treat other conditions, including Alzheimer's disease and schizophrenia. If these approaches bear fruit, lithium could become the 'aspirin of the brain' — just as aspirin is used to combat conditions from arthritis to heart attacks, lithium could treat a range of neurological disorders.

Wonder drug

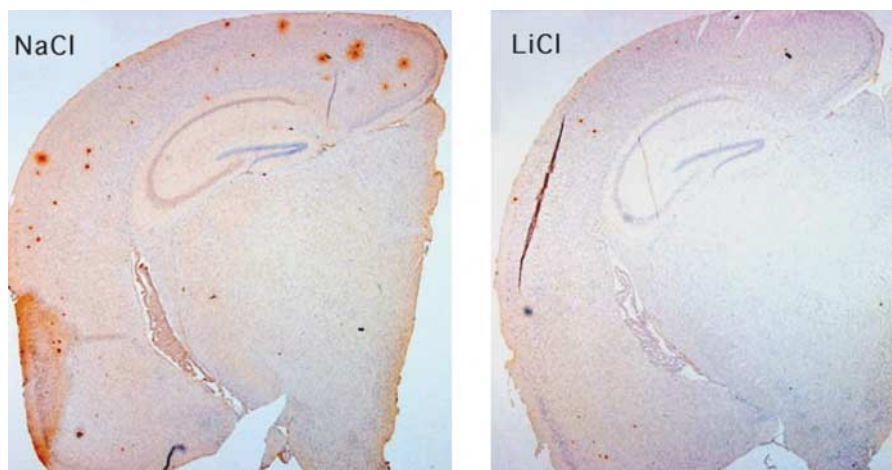
The drug's first advocate was John Cade, senior medical officer at the Victorian Department of Mental Hygiene in Melbourne, Australia. In 1949, Cade reported that lithium had a calming influence on guinea pigs, and that it lessened manic symptoms in humans¹. Twenty-one years and many studies later, lithium was licensed in the United States for the treatment of BPD.

BPD is surprisingly common, affecting 1 person in 100. Without some form of therapy, up to 20% of sufferers will commit

suicide². Given these statistics, lithium has proved to be invaluable. It is not the perfect drug by any means: those who take it can experience short-term side effects such as weight gain and hand tremors, and long-term problems include an increased risk of kidney failure. But up to 80% of sufferers have a positive reaction to lithium³, and the suicide rate is reduced by a factor of eight. "Lithium can keep people well for decades," says Richard Day, a psychiatrist at the University of Dundee, UK.

Our understanding of the causes of BPD is limited. Little is known about how the brains of sufferers differ overall from those of non-sufferers, although there is some evidence that people with the condition have smaller brain volume in certain areas⁴ and that lithium treatment can help to correct this⁵. Knowledge about cellular abnormalities associated with BPD is similarly sketchy.

Given this lack of information, ideas about lithium's action tend to be speculative. One possibility, born in the early 1980s, centred on lithium's effect on a sugar called inositol. This sugar is involved in a signalling pathway that regulates the ability of neurons to exchange signals; lowering inositol levels damps down these signals.



Build-up of protein thought to cause Alzheimer's (orange spots) is cut in brains treated with lithium.



Two faced: treatment of frog embryos with lithium can result in two-headed tadpoles, similar to this. Lithium may target an enzyme involved in many cellular processes.

Lithium inhibits inositol production⁶, so some researchers reasoned that BPD might be linked to excessive signalling by over-active neurons, and that lithium worked by calming these cells.

Mood change

Patients with BPD who take lithium do indeed have lowered inositol levels, but the link with the drug's therapeutic effects is not clear. In 1999, Hussein Manji, a lithium researcher at the National Institute of Mental Health (NIMH) in Bethesda, Maryland, found that lithium causes brain inositol levels to drop. Together with Gregory Moore, a physicist at the Wayne State University School of Medicine in Detroit, Manji found that after a week of lithium treatment, inositol levels in particular parts of the brain fell by 30% in patients with BPD — but their mood did not change for another two to three weeks⁷. "This makes it very difficult to ascribe a therapeutic relevance," says Manji.

In the 1990s, a two-headed tadpole and a developmentally challenged slime mould heralded the arrival of an alternative theory. Researchers knew that the number of spores

produced by the slime mould *Dictyostelium* is cut when lithium is applied⁸ — a reduction that is also seen when the activity of glycogen synthase kinase-3 (GSK-3), an enzyme that helps control several cellular processes, is disrupted⁹. Lithium also prompted frog embryos to turn into two-headed tadpoles¹⁰, and it was believed that this abnormality might be due to disruption in GSK-3 activity.

In 1996, Peter Klein, a developmental biologist at the University of Pennsylvania in Philadelphia, proposed that the effects of lithium on frog embryos and *Dictyostelium* might be due to some kind of effect on GSK-3. "I had an epiphany," he says. He showed that lithium inhibits GSK-3 (ref. 11), and subsequent work has shored up his initial findings. Researchers now know that lithium binds to GSK-3 and causes the activity of the enzyme to drop. Phosphate molecules then attach to GSK-3, deactivating the enzyme further¹². In turn, this affects the expression of the various genes influenced by GSK-3 levels.

But how does this relate to BPD? Some evidence comes from lithium's ability to protect and regenerate cells. Patients with

BPD have up to 40% less grey matter — tissue that is mainly composed of cell bodies — than normal in brain regions associated with mood⁴. Lithium may help counteract this cell loss: a month's treatment boosts the volume of total grey matter by an average of 3% (ref. 5). The drug may do this by prompting the production of new neurons, a process that some researchers believe involves GSK-3. De-Maw Chuang, a neurobiologist in the Mood and Anxiety Disorders Program at the NIMH, has shown that lithium stimulates neural stem cells in culture, causing them to multiply faster¹³.

Such results do more than suggest a connection between lithium, BPD and GSK-3 — they also suggest ways in which the drug could treat other neurological disorders. Chuang, for example, has looked at whether lithium could be used to tackle Huntington's disease, a condition involving neuron loss in the striatum, a region associated with the control of movement. As part of a group of National Institutes of Health researchers, he tracked the effects of lithium on stem cells from a rat model that simulates Huntington's disease.

Neuronal stem cells line the edges of ventricles, cavities containing a fluid that protects the brain. Chuang's team, which has not yet published its results, found that lithium prompts these cells to leave the ventricles and migrate to the striatum. "You see many dividing cells near the site of injury," he says. After lithium treatment, cell death is cut by 80%. "It's too early to say if the same thing is happening in the human brain," Chuang says, "but it is a possibility."

Lithium also helps rats to regain balance and mobility after suffering strokes, in which loss of blood supply to the brain causes areas of neurons to die. Lithium treatment decreased the size of the dead areas by up to 40% (ref. 14). This recovery may, in part, be the result of stem cells dividing to produce replacement cells, speculates Chuang.

Shattered illusions

There is, however, an alternative explanation for lithium's ability to increase brain volume in patients with BPD — and it also has implications for the treatment of other diseases. In 2000, Manji showed that the brains of patients with BPD contained fewer shrunken and withered neurons when they were treated with lithium. He also showed that lithium boosts levels of *N*-acetyl aspartate⁵, a marker for healthy cells in grey matter. This suggests that the drug could be protecting cells in patients with BPD, instead of, or as well as, prompting the growth of new neurons.

Such protective properties suggest that lithium could be used to treat psychotic disorders such as schizophrenia. This

condition, which is characterized by hallucinations and delusions, has been linked in some patients to cell loss. Abnormal levels of some neurotransmitters, the chemicals used to send signals between neurons, are thought to be involved. High concentrations of one neurotransmitter — glutamate — can kill cells.

In cell culture, lithium can protect cells against glutamate-induced death¹⁵, although it's unclear how it does this, and whether GSK-3 is involved. Researchers are investigating several leads, including the possibility that lithium turns on genes that help to protect cells against glutamate or prompts cells to release nourishing growth factors.

Whatever the mechanism, the new studies are enough for scientists to want to test lithium in humans with psychosis. "Given early on, a low dose of lithium may delay or even prevent the onset of psychotic disorders, including schizophrenia," says Gregor Berger, a psychiatrist at the University of Melbourne's Personal Assessment and Crisis Evaluation Clinic. He is testing a group of 30 young adults deemed to be at risk from developing psychosis — they have a family history of schizophrenia or a record of brief psychotic disturbances. Without treatment, about a third would be expected to develop psychosis¹⁶. One year on, his unpublished study shows that all of the subjects given daily doses of lithium are free from symptoms.

Untangling the knot

Lithium might also ward off Alzheimer's disease, a form of dementia. This disease has two hallmarks in the brain: protein plaques and tangles of protein fibres. Tangles are thought to be caused by the addition of phosphate molecules to a protein called tau, a process that involves GSK-3 (ref. 17). This might explain why, in cultured cells at least, lithium can prevent tau phosphorylation¹⁸.

The drug might also tackle the plaques seen in Alzheimer's disease. These insoluble deposits of the protein amyloid build up around neurons, impairing the cells' ability to communicate. When GSK-3 is added to cultured cells, the levels of amyloid protein go up. In mice that have been genetically engineered so that they are prone to plaques, lithium inhibits GSK-3, stops the protein from building up¹⁹ and cuts the number of plaques. Human trials have not yet been conducted, but Klein suggests that, given early on, lithium might reduce the rate of tangle and plaque formation, or prevent them altogether, ultimately putting the brakes on dementia.

So, are drug companies rushing to set up trials for the treatment of Alzheimer's disease and other conditions? Unfortunately not. The original patent on the use of lithium as a drug has long since expired. It cannot be



Brain power: De-Maw Chuang is probing links between lithium and neural stem cells.

repatented, and so lithium lacks the potential for profit. "Drug companies prefer to plough their resources into more profitable alternatives," says Berger. This makes his research, which involves human subjects, something of a rarity. Berger's research is currently funded by the Stanley Medical Research Institute, a Bethesda-based non-profit organization that supports research on schizophrenia and BPD.

Drug companies and doctors are also deterred by lithium's side effects. In younger people, these can be managed. Whether or not elderly people could tolerate the drug is a separate issue. With age, the problems can worsen, cautions Day. "A fine tremor could become a coarse tremor," he says. Nausea could become vomiting, and, in extreme cases, drowsiness might slip into coma.

Pharmaceutical companies are focusing instead on developing lithium-like drugs that have less extreme side effects. GSK-3 occurs in two forms, known as α and β , but only GSK-3 α is required for amyloid production. Klein speculates that molecules that selectively

inhibit the α version might rid the brain of plaques while minimizing adverse reactions. Most major pharmaceutical companies are currently developing GSK-3 inhibitors, but are tight-lipped about the details.

If such research produces a better alternative to lithium, no one will complain. But in the meantime, lithium's reputation is being tarnished. The drug has a narrow therapeutic window — too little is ineffective, too much is toxic — so finding the right dose can take time. Even then, patients need regular psychiatric assessments and blood tests. Such problems mean lithium is losing popularity as a treatment for BPD, just when its activity is beginning to be understood. "Most doctors don't know how to use it and are needlessly afraid of lithium," says Martha Hellander, executive director of the Child & Adolescent Bipolar Foundation, Wilmette, Illinois. "Lithium has a tainted reputation," agrees Manji.

This, say advocates of the drug, is restricting its use. Hellander would like to see it studied in children with BPD. Currently, although the condition can be diagnosed before puberty, no medications are approved for the treatment of children under 12 years old with BPD in the United States.

And lithium could still find a use in Alzheimer's disease, despite its side effects. Lower doses would cut the side effects, yet could still be beneficial, says Berger, who is using low doses in his study. But unless lithium gets some better press, and its action becomes better understood, such potential benefits could go unexploited. "Hardly anyone speaks up for lithium," says Hellander. "It's like the orphan child of treatment."

Helen R. Pilcher works in Nature's news syndication team.

1. Cade, J. F. *J. Med. J. Aust.* **2**, 349–352 (1949).
2. Tondo, L., Hennen, J. & Baldessarini, J. J. *Acta Psychiatr. Scand.* **104**, 163–172 (2001).
3. Price, L. H. & Heninger, G. R. *N. Engl. J. Med.* **331**, 591–598 (1994).
4. Drevets, W. C. *et al. Nature* **386**, 824–827 (1997).
5. Moore, G. J., Bechuk, J. M., Wilds, I. B., Chen, G. & Manji, H. K. *Lancet* **356**, 1241–1242 (2000).
6. Berridge, M. J., Downes, C. P. & Hanley, M. R. *Cell* **59**, 411–419 (1989).
7. Moore, G. J. *et al. Am. J. Psychiatry* **156**, 1902–1908 (1999).
8. Maeda, Y. *Dev. Growth Differ.* **12**, 217–227 (1970).
9. Harwood, A. J., Plyte, S. E., Woodgett, J., Strutt, H. & Kay, R. R. *Cell* **80**, 139–148 (1995).
10. Kao, K. R., Masui, Y. & Elinson, R. P. *Nature* **322**, 371–373 (1986).
11. Klein, P. S. & Melton, D. A. *Proc. Natl Acad. Sci. USA* **93**, 8455–8459 (1996).
12. Jope, R. S. *Trends Pharmacol. Sci.* (in the press).
13. Hashimoto, R., Senatorov, V., Kanai, H., Leeds, P. & Chuang, D.-M. *Neuroscience* **117**, 55–61 (2003).
14. Ren, M., Senatorov, V. V., Chen, R. W. & Chuang, D. M. *Proc. Natl Acad. Sci. USA* **100**, 6210–6215 (2003).
15. Hashimoto, R., Hough, C., Nakazawa, T., Yamamoto, T. & Chuang, D. M. *J. Neurochem.* **80**, 589–597 (2002).
16. McGorry, P. D. *et al. Arch. Gen. Psychiatry* **59**, 921–928 (2002).
17. Sperber, B. R., Leight, S., Goedert, M. & Lee, V. M. *Neurosci. Lett.* **197**, 149–153 (1995).
18. Stambolic, V. *et al. Curr. Biol.* **6**, 1664–1668 (1996).
19. Phiel, C. J., Wilson, C. A., Lee, V. M. & Klein, P. S. *Nature* **423**, 435–439 (2003).



Lithium could help prevent the cell loss that impairs the movement of Huntington's sufferers.

Science and ethics must not be separated

The progress of research must be kept free from religious and political intervention.

Sir — It is regrettable that ethics has been split from science and renamed bioethics. Ethics is an integral part of science. Like science, it requires us to be consistent and empirically justified in our interpretations of the actions of scientists. The ethics of science and science itself share the goal of comprehending in human terms scientists' actions in manipulating the physical world.

The division of science and ethics has been driven by an increasing interest in the actions of scientists by non-scientists. An unfortunate result of this has been a shift away from the consistent and justifiable methodology of science to an approach based on an often ill-defined 'personal philosophy' and 'gut feeling', for instance as described in Correspondence by D. P. Leader: "Reproductive cloning: an attack on human dignity" (*Nature* **424**, 14; 2003).

Such 'gut feelings' undoubtedly play a part in science but are useless for proper understanding. The reactions of non-specialist observers to complex ethical problems raised by cutting-edge science such as embryonic stem-cell research are no more justified or useful than their opinions about the technical difficulties yet to be overcome. The central issue in the ethical debate surrounding the embryo is not whether it is wrong to kill innocent human beings, but what the embryo and its disaggregation constitutes. The specialist scientific community's familiarity with the facts places it in a privileged position in determining the interpretation or interpretations best supported by the facts.

The rise of bioethics as an independent discipline has resulted in a confrontation between ethics and science that has

obscured the similar aims of both. Of considerable concern is the increasingly political and religious nature of bioethics and the power it wields over the direction and progress of science. Science and the ethics of science are two sides of the same coin, dealing with the same empirical data and actions of the same scientists.

As well as thinking of their actions in terms of future experimental design, scientists must explain the significance of their actions in the wider scientific and human contexts. Scientists must take the lead in ensuring that the progress of science is both ethical and as free from political intervention as possible, if for no other reason than that only they can do so.

Paul Copland

Department of Biochemistry, University of Otago, Box 56, Dunedin, New Zealand

Safe home planned for Iowa herbarium

Sir — The plight of natural history collections is regrettable, as you report in the News story "Natural history collections in crisis as funding is slashed" (*Nature* **423**, 575; 2003). You cite some examples of such collections in jeopardy, including the University of Iowa herbarium collection which is being moved to Iowa State University. In an era when state budgets are shrinking and debt is increasing, we feel that this cooperative venture is the most responsible way of managing and caring for a resource unique to the state.

The collections are not in jeopardy, as you suggest, but are being consolidated. As the state land-grant university, Iowa State University maintains active programmes in plant sciences, including an herbarium, while the University of Iowa has emphasized plant molecular biology.

The University of Iowa herbarium has not been a major research/teaching tool for many years. During 1998–2002, work at the Iowa State herbarium was cited in 75 publications, while work at the University of Iowa herbarium resulted in some 18 citations. Similarly, only two courses now use the herbarium at the University of Iowa, whereas herbarium use has remained steady at Iowa State. To ensure that the few faculty members at the University of Iowa using the resource have continued access, we will support travel between the two universities and have agreed that specimens needed for ongoing research may be kept on long-term loan.

Our intention is to retain all specimens

that are not duplicated; the Iowa State herbarium has already applied for funds to accommodate the merged collections. If it becomes necessary to disperse some specimens, we will relocate them at institutions where they will be maintained and cared for. We began this process in the mid-1980s, when the bulk of the University of Iowa fungal collection was transferred to Iowa State University. Even if Iowa State cannot accommodate the entire remaining University of Iowa herbarium, it will keep all material collected in Iowa.

Jack Lilien

Department of Biological Sciences, University of Iowa, Iowa City, Iowa 52242-1324, USA

Computer's ability to verify proof is an illusion

Sir — Your news feature "Does the proof stack up?" (*Nature* **424**, 12–13; 2003) — addressing the difficulties faced by mathematicians in verifying the computer-aided putative proof of Kepler's conjecture concerning the densest arrangement of spheres — contains the seemingly reasonable statement by one mathematician: "I believe in a proof if I understand it". Yet it is well known that many a supposed understanding is a misunderstanding, and that there is no general rule guaranteed to distinguish between these opposite possibilities.

Arguments over the validity of mathematical proofs have raged for as long as mathematicians have been trying to prove conjectures. The history of this formal science is strewn with 'proofs' that

have been accepted and later rejected, and 'proofs' that have been simultaneously accepted and rejected by different mathematicians.

The understanding of a computer program is no more reliable than the understanding of a mathematical proof.

Computers lure us into unmanageable complexity. It is, after all, one of their great assets that we can construct very long sequences of simple manipulations for a computer to execute at high speed and with unerring accuracy. However, this complexity (as with a list of alternative cases in a computer-aided proof) soon compounds, such that 'understanding' the program is impossible and a proof of the program is as complex as the program itself, or even more complex.

Scientists seek truths about the world (or about formal systems that mirror the world) and they do their best to confirm the validity of any new truth uncovered. But the reality is that (apart from certain limited totally abstract constructions) scientific truth can never be secure, validity can only be up to a certain point and formal verification is not a practical option. Scientific truths are best-guesses given all the available evidence, or, more pessimistically, they are potential errors as yet unmasked. In sum, computers add nothing new to the difficulties of proof and verification — they simply enlarge existing problems through the magnifying glass of complexity escalation.

Derek Partridge

Department of Computer Science, Harrison Building, University of Exeter, Exeter EX4 4QF, UK

An ethical dilemma

The rise and fall of UmanGenomics — the model biotech company?

Hilary Rose

For the biotechnology industry, data that cover the health and genetics of a population can represent a treasure-trove of information. But the marriage of business and such public data is, at best, uneasy. Conflicts over informed consent and privacy have, for example, overshadowed plans by Reykjavik-based deCODE Genetics to establish a health-sector database covering most of Iceland's population. So it was with great fanfare a few years ago that the leading scientific press welcomed a new Swedish biotechnology company: UmanGenomics^{1,2}.

Promising that its foundations would be based on popular consent, and pledging an ethical approach to the data it would access, including the public sharing of results, UmanGenomics was promoted as “a model of how public tissue banks should interact with the biotech industry”². The biobank in question was created during a health-intervention campaign in the northern Swedish county of Västerbotten. During the project, clinical researchers at Umeå University collected and collated lifestyle and health data, including blood samples, for a significant proportion of the county's population.

UmanGenomics grew out of the shared wish of the university and Västerbotten's county council to commercialize the biobank that had been created under their auspices. The original plan was to “discover disease-related genes, explore their function and market this knowledge”. But just four years later, the company has been more-or-less moth-balled. No contracts have been secured, and no revenue was achieved during 2001. At the end of 2002 there were just 16 full-time staff — by May 2003 this was down to two.

Meanwhile, a furious row has broken out, primarily between Umeå University as the lead owner of the company and Göran Hallmans, the originator and initial director of the biobank. Not only has the conflict been played out in two local newspapers, *Västerbottens-Kuriren* and *Västerbottens Folkblad*, with academics and even national politicians intervening, it has also been followed closely by a multidisciplinary research team led by bioethicist Mats Hansson of Uppsala University, which has published two reports^{3,4}.

A major stumbling block in the development of UmanGenomics was intellectual property rights. Unlike many countries, Swedish law allows for ‘the teacher's exception’, which gives academics ownership of the intellectual property that they produce. But legally it remains unclear how this exception relates to



Valuable resource: Umeå's hospital holds a tissue bank with samples and data from some 87,000 people.

biobanks. The contract drawn up to allow UmanGenomics access to the Västerbotten biobank recognizes only the county and the university as “principals with rights of disposal” over the data the bank holds. Hallmans, as the chief architect behind the biobank's creation, is challenging this interpretation.

The conflict has become bitter and there is little room for compromise. The university and the county see Hallmans as an impediment to their plans for scientific progress and high-tech industrial development in Umeå. Hallmans, supported by current and retired colleagues, is outraged that the contract between the county and the university has allegedly set aside pre-existing research contracts; that UmanGenomics (like deCODE in Iceland) has been granted monopoly commercial access to the biobank; that the contract ignores the fact that donors did not sign up to have their samples exploited for private profit; and that his intellectual property rights have been ignored. In October 2001, the first

legal complaints by biobank colleagues of Hallmans were rejected and the case referred to the appeal court, where a judgement is still pending. For a company founded so determinedly on ethics, where did it all go wrong?

Building a biobank

By 1998, it was obvious that the biobank assembled by Hallmans and his colleagues, and held at the county hospital in Umeå, was a potential gold-mine for genomics. To capitalize on this amassed information, Umeå University and the county council of Västerbotten decided to create a company that could make commercial use of the data. To those ends, Sune Rosell, then recently retired from drug firm Astra where he had sometime been research director, was appointed as an adviser and chair of the biobank. In the light of his advice, the bank and the commercial interests were split. Rosell was appointed chief executive of the newly created UmanGenomics, and the biobank remained the university's and county's responsibility.

Rosell's plans for UmanGenomics were developed within the framework of the ethical guidelines for genetics drawn up by the Swedish Medical Research Council and the National Board of Health and Welfare's proposals for legislation on biobanks. In turn, the government made it clear that any such legislation would need to be compatible with the 1997 European Convention on Human Rights and Biomedicine to simplify eventual ratification. By contrast with deCODE Genetics, which claimed the genetic uniqueness of the Icelandic population as the key selling point for its biobank project, Rosell



Göran Hallmans: helped create Umeå's biobank.

opted to place ethics at the heart of the sales pitch for UmanGenomics.

But UmanGenomics soon ran into difficulties. One early hiccup was the plan for a public-private partnership to own the company. This provided for the university and the county — as the perceived owners of the biobank — to hold 51% of the shares and for the rest to be privately held. But in Sweden, public bodies cannot engage in commercial activity. Fortunately, universities in Sweden, as in many other countries, are encouraged to stimulate industrial-academic links and ventures. So to maintain the ethical-ownership model for UmanGenomics, the university became the sole public shareholder.

A second hitch was the marketing plan. The company hoped to secure contracts with the big ten drug companies that had funded the SNP Consortium, an initiative to build a publicly accessible database of human single-nucleotide polymorphisms (SNPs). But in the event, no such contracts were secured.

Third, UmanGenomics had to switch its business plan from gene hunting to the newer, hotter areas of functional genomics and proteomics. Reorienting was costly and the company argued that it would need an additional 200 million Swedish kronor (US\$23.5 million) to achieve its goal. To secure this fresh investment it claimed that the 51% public ownership was unhelpful and that potential investors were being scared off by the requirement that results had to be available to competitors. Ethics, which had first been the selling point of UmanGenomics, now threatened to drive the firm into bankruptcy. Fortunately for the company, the Swedish biobanks law introduced at the beginning of 2003 made it possible to argue that the 51% public ownership was no longer needed, and it was agreed to negotiate a new contract.

A valuable collection

These difficulties were exacerbated by the concurrent fight over access to the biobank. This problem was inherent to the original commercialization plans, which were developed as if the bank was simply the result of the county's health-promotion activities beginning in the mid-1980s. Yet the objective of the project was to reduce the levels of heart disease and premature death in the county. All that was needed for this was patient-based records of lifestyle, health status and such measures as lipid levels and blood pressure. Keeping blood samples was not intrinsic to the project.

There are two issues that no one seems to have taken fully on board at the outset. Who had found the resources to run the biobank over its 15-year existence? And what role should those who established the bank play in the new venture? It is acknowledged by all of the players in this story that the biobank was created by a number of clinical researchers, with Hallmans as the key figure. Biobank research was unfashionable until the mid-



On hold: UmanGenomics has been mothballed.

1990s when genomics came onto the agenda, and so did not easily attract funds. But Hallmans secured several external grants, developed international collaborations and used projects for unemployed people to train a workforce to take care of the bank.

As the potential for genomics research grew, Hallmans made an initial, unsuccessful attempt in the early 1990s to use the biobank for commercial research. Later, together with his colleague Joakim Dillner, he proposed a social-enterprise model in which ownership would be vested in a foundation representing the people of Västerbotten. This was to be managed by the researchers, the county and the university, as well as by potential funders such as the Knut and Alice Wallenberg Foundation. All profits were to be reinvested into biobank research. Hallmans' group saw this as meeting the trust placed in them by the donors, benefiting research and honouring the bank's commitments to its existing funders. But the plans drawn up by the university and the county cut across this scheme.

As late as February 2003, Umeå University's ethics committee, in its invited comment on the draft contract, was still insisting that the biobank's position in relation to both the county and the university was unclear and should be investigated. Further, it argued that the proposal in the contract to appoint a steering committee of just four people to set the criteria for research using the bank's samples breached both the committee's and the bank's responsibilities under the Declaration of Helsinki on medical research involving human subjects. None the less, a month or two later, Hallmans had been removed from his post as director of the biobank.

Unless the court of appeal can find a solution that will satisfy both sides, this is a lose-lose story. There will be no dynamic biotechnology company with an ethical relationship to a public biobank for a small Swedish city that could use some industrial diversification. Instead there will be, at best, a 'virtual' biotech company.

Could it have been different? This account

draws attention to the difficulty of exploiting an existing public biobank through a start-up biotech company without the involvement of all of the relevant stakeholders — not least the researchers who established the bank. This point reinforces a conclusion from commercial lawyers Urban Paulsson and Rebecka Frisk, who observe: "Because it is uncertain in Swedish law whether the biological research results initially belong to the scientist or to the university, our advice to an outside contracting party ... is to ensure that ownership is assigned to both," (ref. 4, p. 261).

Alternative approach

The preoccupation with 'ethics' at UmanGenomics seems to have led to the neglect of both the researchers' potential property claims and the company's needs for flexibility in a dynamic financial and scientific context. Might the Hallmans-Dillner social-entrepreneurial model have fared better? Arguably, such an approach would more likely have secured the support of all the stakeholders, from donors to researchers to the responsible authorities.

Less glitzy than the admired ethical model, a social-entrepreneurial approach might have been able to develop slowly and steadily from the existing research base using both public and foundation money. This could have been helpful after the NASDAQ crash, which left venture capital chary of biotechnology. Non-commercial status, broadly echoing that of Britain's Biobank UK — with the difference that Biobank Umeå would have been very locally rooted — could have had advantages.

The idea of bottom-up growth has some scientific attractions. Although there is considerable enthusiasm in some quarters for Biobank UK, there are also some within the scientific community who are concerned not only about whether the costs of setting up the bank will take too large a slice of the research budget, but whether the bank will be adequately funded regardless. A bottom-up project could bypass this problem, growing within its strength, building on the existing biobank resource, and expanding flexibly to meet scientific and commercial opportunities as they arise. In addition, a social-entrepreneurial model would have enabled biobank research to be attentive to the interests of all, and not just some, of the stakeholders. As Umeå — company, county and university — has sadly learnt, failing to keep all the stakeholders on board is expensive. ■

Hilary Rose is a visiting research professor in sociology at City University, 4 Lloyd Square, London WC1X 9BA, UK.

e-mail: hilaryrose@blueyonder.co.uk

- Abbott, A. *Nature* **400**, 3 (1999).
- Nilsson, A. & Rose, J. *Science* **286**, 894 (1999).
- Hansson, M. G. (ed.) *The Use of Human Biobanks* (Uppsala Univ., Uppsala, 2001).
- Hansson, M. G. & Levin, M. (eds) *Biobanks as Resources for Health* (Uppsala Univ., Uppsala, 2003).

♦ www.umangenomics.com

Net losses

Overfishing in the north Atlantic Ocean has left fish stocks in peril.

In a Perfect Ocean

by Daniel Pauly & Jay Maclean

Island Press: 2003. 175 pp. \$50, £38.50 (hbk); \$26, £19.50 (pbk)

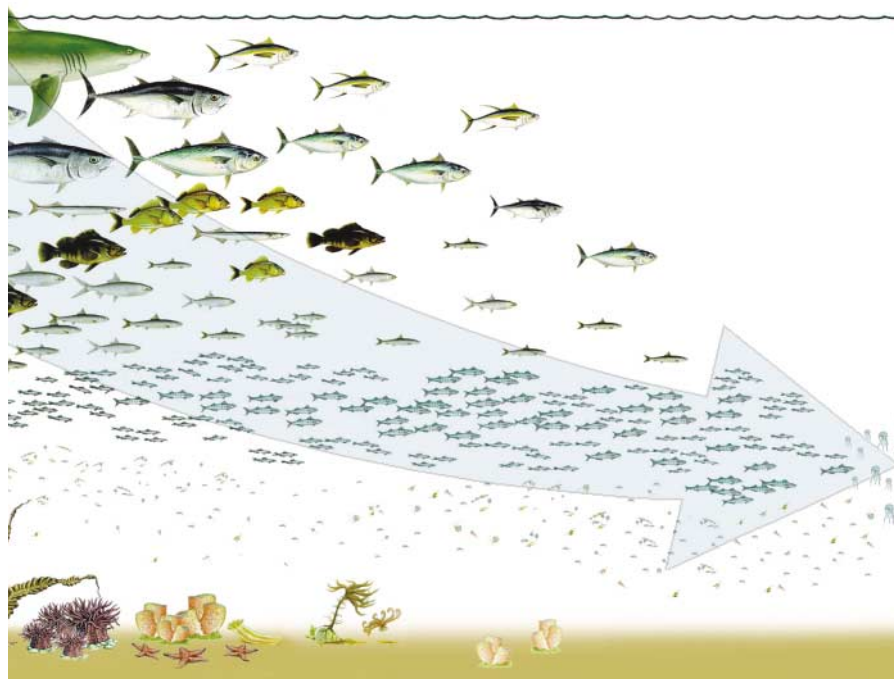
John Shepherd

Daniel Pauly is a pioneer of bold new methods for the assessment of marine ecosystems. His methods for estimating natural mortality rates, for interpreting length-frequency distributions and for quantifying energy pathways in marine food webs were all highly controversial when they were introduced, but are now accepted as providing useful rough and ready approximations to reality. In this new book, Pauly and co-author Jay Maclean report the results of another, even bolder, venture: an attempt to undertake a holistic assessment of the state of fisheries and ecosystems in the whole of the north Atlantic Ocean.

This work has been carried out as part of the "Sea Around Us Project" (see www.saup.fisheries.ubc.ca), a programme based at the University of British Columbia that involves an international team of researchers and reviewers. *In a Perfect Ocean* provides an account of this work for a general audience, to supplement several publications in the professional literature. It includes a good historical review of the marine resources (especially the fish stocks) of the north Atlantic, but the centrepiece is the analysis of their decline as a result of exploitation, mostly during the twentieth century.

Pauly and his team have used his Ecopath with Ecosim software (www.ecopath.org) to quantify food webs and trophic pyramids, along with some statistical extrapolations to areas where the data are inadequate, to create an overview for the whole ocean basin. This will probably be controversial, but my overall impression is that they are unlikely to be far wrong. After all, much of the story, for such areas as the Grand Banks off Newfoundland and the North Sea, is already well documented and depressingly familiar.

The essence of their story is contained in two striking series of maps (figures 8 and 14), which show the spatial distributions of the rise in fishing intensity and the decline of fish biomass from 1900 to 1999. We have known for years that fishing intensity (expressed as a catch/biomass ratio) has been close to 1.0 in many of the most productive areas for several decades. Pauly and Maclean's holistic view now reveals a resultant decline of biomass, by a factor of ten or more, in those same areas. Similar estimates have been available for selected stocks for many years through international organizations such



Fishing down the food web: when large fish are gone, small ones further down the food web are caught.

as the International Council for the Exploration of the Sea (www.ices.dk) and the Northwest Atlantic Fisheries Organization (www.nafo.ca). But these estimates were not taken too seriously even by the scientists whose models produced them, because they might have taken insufficient account of associated changes in predator and prey abundances, and thus in natural mortality and growth rates. However, recent work based on an analysis of historical survey data (reported by R. A. Myers and B. Worm, *Nature* **423**, 280–283; 2003) provides strong support for their credibility.

This analysis by Pauly and colleagues, which specifically includes the effects of interactions in the trophic web, provides further support for the reality of massive declines of heavily exploited fish stocks. We can no longer comfort ourselves with the belief that nature will readjust and ameliorate the adverse effects that we have wrought.

The book contains much else of interest, including a good account of how we got into this mess, and a much less satisfactory chapter on what to do about it. Like everyone else, the authors conclude that substantial reductions in fishing effort are required, but they have little to add to the existing portfolio of tools (catch quotas, closed areas and vessel buy-back schemes) for achieving this.

In their preface, Pauly and Maclean remark that the scientific advice on managing sustainable fisheries has mostly gone

unheeded. Indeed it has, and the reason for this is simple: all effective conservation measures require that fewer fish be killed, and that means (in the short term) that fishermen's incomes will be reduced. This creates a fundamental conflict between short-term losses and long-term gains that has not yet been resolved anywhere in the world. Unless and until we can find a way to reconcile these, we must expect fishermen and their representatives to oppose with vigour anything that might actually work.

One possible way forward could be the use of short-term transitional aid, such as an effort buy-back scheme (to be regarded as an investment in a better future) that could be paid for in the long term by substantial resource rental charges. However, Pauly and Maclean do not really address the powerful economic forces that have prevented progress for so long, nor possible ways in which the system might be engineered to make them less destructive. This is a matter of regret, because most of the methods they discuss have already been tried, and have not yet solved the problem.

In particular, they suggest a carbon tax (affecting fuel costs) as a potential solution. But on their own figures for tons of fuel used to catch tons of fish, a carbon tax of about US\$50 per ton of fuel is unlikely to affect the behaviour of fishermen very much, if it allows them to land fish worth approximately \$5,000. Something

more subtle and effective is needed.

Overall, this book is a useful contribution to the literature, and one can hardly disagree with the conclusion that overfishing is “a primary cause of ecosystem disruption”. It has some excellent images; in particular figure 17 (shown overleaf), which is a graphic illustration of what it means to fish down the food web, and deserves to be widely reproduced. But I found the book’s structure awkward: it has no less than 30 pages of prefaces, and another 30 pages of endnotes, including some substantial discussions of important issues (such as whaling and international institutions), running to several pages each.

Those who work in the field will find that the book is a bold attempt to create an ocean-wide overview that complements the more conventional stock-by-stock reductionist methods. General readers will find a broad, and in places passionate, account of the state of a whole ocean and its resources that is both accurate and informative. One can only hope that it will help to motivate more strenuous and effective efforts to deal with the problems so clearly identified, which is, I am sure, what its authors intend. ■

John Shepherd is at the Southampton Oceanography Centre, Empress Dock, University of Southampton, Southampton SO14 3ZH, UK.

More on fisheries

Handbook of Fish Biology and Fisheries, Vols 1 and 2

edited by Paul J. B. Hart & John D. Reynolds
Blackwell, £130

New in paperback

I Have Landed

Stephen Jay Gould
Vintage, £7.99

The Constants of Nature

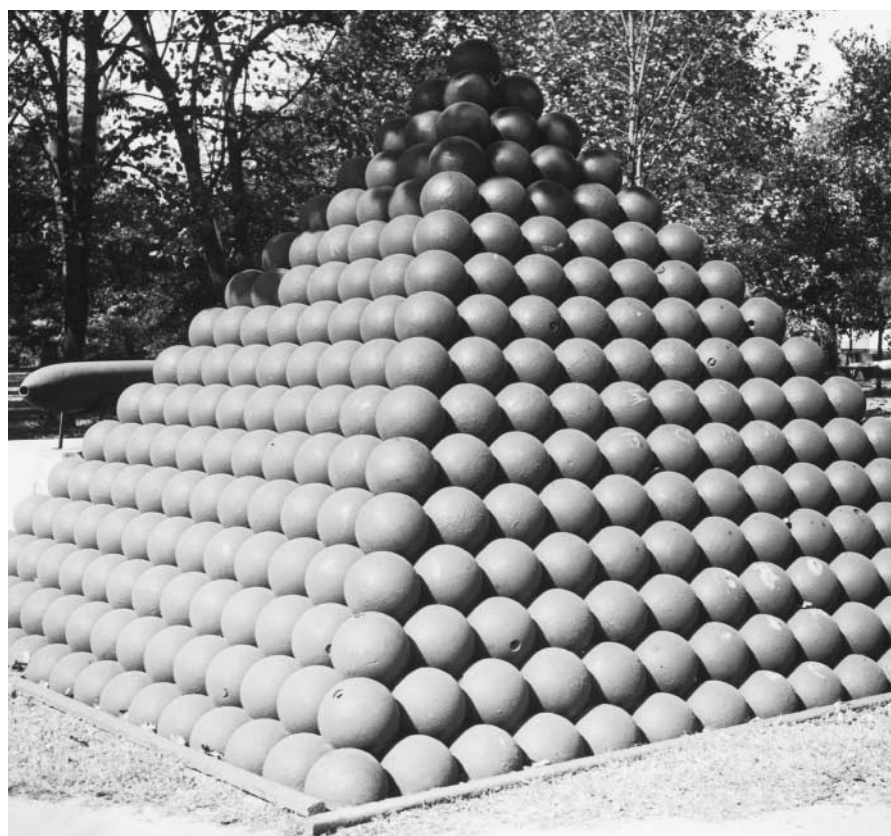
by John Barrow
Vintage, £8.00
“Barrow discusses the role of constants of nature, the historical quest to understand them, the role of the anthropic principle as a guiding philosophy and some recent evidence suggesting that some of the constants of nature are probably not constants at all.”
Thanu Padmanabhan *Nature* **419**, 780 (2002).

The Borderlands of Science: Where Sense Meets Nonsense

by Michael Shermer
Oxford University Press, £9.99

The Way of the Cell: Molecules, Organisms and the Order of Life

by Harold M. Franklin
Oxford University Press, £12.95



Packing them in: cannon balls stored in a pyramid are stacked together as densely as possible.

The proof of the packing

Kepler's Conjecture: How Some of the Greatest Minds in History Helped Solve One of the Oldest Math Problems in the World

by George G. Szpiro
Wiley: 2003. 304 pp. £18.50, \$24.95, €24.95
Neil Sloane

The classical sphere-packing problem is to determine how densely a large number of identical spheres (such as ball-bearings) can be packed together in a finite space. In 1611 the German astronomer Johannes Kepler stated that no packing could be denser than that of the face-centred cubic (f.c.c.) lattice arrangement favoured by grocers for stacking oranges, which fills about 0.7405 of the available space. It took mathematicians some 400 years to prove him right.

Kepler's Conjecture gives an entertaining and readable account of the history of the problem and the attempts to solve it, culminating with Thomas Hales' successful proof, announced in 1998. George Szpiro also discusses a large number of peripherally related topics, including David Hilbert's list of 23 unsolved mathematical problems from 1900 (Kepler's conjecture is part of problem 18), the kissing-number problem (how many balls can touch another ball of the same size),

linear programming and Lord Kelvin's soap-film problem.

The book is a mixture of mathematics, history and anecdotes. In his research, the author has found many good stories to retell. Even people familiar with the subject will find new anecdotes here, and it seems that most of them are more-or-less true, although one might quibble with the details. Did John Conway's father really teach chemistry to two of the Beatles? Well, sort of.

The tone of the remarks is sometimes derisive, which some readers may find offensive rather than humorous. Young Carl Friedrich Gauss is described as a “little squirt”, ‘wrangler’ is “one of those esoteric blue-ribbon signs of esteem... reserved for British overachievers”, and sheaf theory is a “major bore”. And after a rather harsh discussion of the attempts of the great Hungarian geometer László Fejes Tóth (his name is consistently misspelled in the book) to prove the dodecahedral conjecture, Szpiro writes: “One might come away from this chapter with the impression that Fejes-Tóth was a bumbling dreamer whose work mostly contained unfulfilled promises and unproven hypotheses. This does not represent the whole picture.” Indeed not.

The mathematical content is less satisfactory than the historical part. As William Barlow described in *Nature* in 1883 (**29**, 186–188), the f.c.c. packing can be built up by layers. Put down a layer of spheres arranged in a triangular lattice — the arrangement used

when racking billiard balls — and place another layer on top, and repeat. There are two ways to place subsequent layers. Viewed from above, there are three different positions for the centres of the spheres in any one layer, say A, B and C. If the layers follow the order A, B, C, A, B, C, ..., then the f.c.c. packing is obtained. If they follow the order A, B, A, B, A, B, ..., then an equally dense packing known as the hexagonal close packing (h.c.p.) is obtained.

Kepler's conjecture is that there are no packings that are denser than the f.c.c. or the h.c.p. packings (or any one of the infinite number of different packings obtained by varying the order of the layers). The f.c.c. and h.c.p. packings have the same density, but they are different: one is a lattice, the other is not. Spiro claims that the f.c.c. and the h.c.p. are "the exact same packing, viewed from different angles". They are not.

Another distraction in the mathematical discussions (which fortunately are set in a different typeface, so they can — and should — be skipped by the casual reader) is the author's misuse of the word 'surface'. Several times he writes of the surface of an object, when he means its area, or even its volume.

One of the oldest theorems about sphere packing was proved by Gauss in 1831, when he showed that the f.c.c. is the densest lattice packing of spheres. Szpiro attempts to reproduce Gauss's proof, but makes a mess of it. For example, on page 255 the determinant needs to be negated, and denoted by a new symbol, Δ , say. Then six occurrences of the letter D on that page need to be changed to Δ . Similar repairs are needed on the next page.

The book hardly mentions one of the main reasons for studying the packing of spheres: its application to digital communications. From the communication theorist's viewpoint, Hales' result on three-dimensional sphere packing is just the beginning of the story. One of the fundamental questions in communication theory is to determine the densest packing of equal balls in multi-dimensional space. A geometrical way of representing signals, which is at the heart of Claude Shannon's mathematical theory of communication, underlies the high-speed modems that we now take for granted.

Szpiro mentions this subject only briefly, in the final chapter, but the discussion is marred by another error. He describes the following problem as a far-fetched application of packing problems (it is actually a standard type of problem in error-correcting codes). The problem is to find as many strings of ten decimal digits as possible, subject to the constraint that any two of the strings must differ by at least two units in each position. He misuses the known bounds on the density of sphere packing in ten-dimensional space to conclude that "at least 400,000,000 signals can be represented, which is sufficient for all words in all languages of the

globe". However, the correct answer is not 400,000,000, but 5.

One can only admire Szpiro's valiant attempts to explain the different approaches used by Richard Buckminster Fuller, Wu-Yi Hsiang and Hales in their attacks on the problem (although the serious reader would do better to read Hales' own descriptions). Szpiro's discussion of the arguments between the protagonists is certainly entertaining. He illustrates them with a quotation from Henry Kissinger, who "was once asked why departmental fights are so violent, why back-stabbing is so common among academic colleagues. His answer was short and to the point: 'Because the stakes are so small.'" Typically, not quite relevant, but a good story.

As long as readers skip over the technical sections, the book can be recommended as a readable and informative account of a fascinating chapter in the history of geometry.

Neil Sloane is at the AT&T Shannon Laboratory, 180 Park Avenue, Florham Park, New Jersey 07932-0971, USA.

The rise and fall of the Universe

Alpha and Omega: The Search for the Beginning and the End of the Universe

by Charles Seife

Viking Press: 2003. 304 pp. \$24.99

Doubleday: 2003. £18.99

Peter Coles

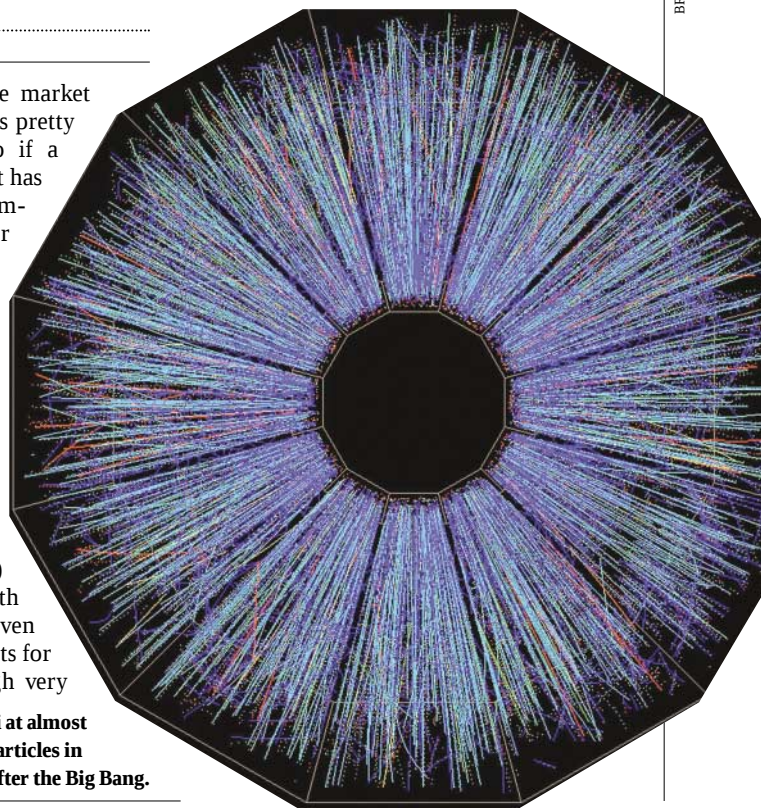
The potentially lucrative market for popular cosmology is pretty crowded these days, so if a book is to be successful it has to stand out from its competitors. One strategy for a publisher is to sign up a professional scientist with something special to say. João Magueijo's *Faster Than the Speed of Light* (reviewed in *Nature* **422**, 563–564; 2003) and Janna Levin's *How the Universe Got its Spots* (Weidenfeld & Nicolson/Princeton University Press, 2002) are two recent books, both written in distinctive, even quirky, styles by specialists for a lay audience. Although very

The collision of gold nuclei at almost the speed of light creates particles in conditions like those just after the Big Bang.

different, these books have much in common. Both are insiders' views of the subject, both are highly original because the subject matter is seen from the perspective of the authors' own research, and both include a lot of autobiographical material.

Few scientists are capable of putting their understanding and experiences into words as effectively as these two, so publishers have instead enlisted professional writers to look at the subject from the outside. A science journalist may not have as deep an understanding of the technicalities as a research scientist, but may be more experienced at writing for the general public and consequently better at getting the basic ideas across. Particularly successful examples of this genre are *The Whole Shebang* by Timothy Ferris (Weidenfeld & Nicolson/Simon & Schuster, 1997) and, more recently, Bill Bryson's *A Short History of Nearly Everything* (reviewed in *Nature* **424**, 725; 2003), which both demonstrate that winners need not necessarily be on the inside track. Sadly, *Alpha and Omega* by Charles Seife is not among the medal positions.

The book starts promisingly enough, if you can forgive the pseudo-religious overtones of the title (a reference to the Book of Revelations). The suggested emphasis on both the beginning and the end seems a good idea, as there are many books about the birth of the Universe but relatively few about its death. Unfortunately, despite the claims made on the jacket, this theme isn't really taken up by the book itself, except for a few comments in the final chapter.



Instead we have a fairly conventional account of the historical development of cosmology from antiquity to modern times. This account is up-to-date, including such developments as the preliminary release of data from the Wilkinson Microwave Anisotropy Probe and the latest observations of distant supernovae, and is accompanied by some nice illustrations. It is, for the most part, quite well written, but there is too much repetition, some of the diagrams are incomprehensible, and the text is peppered with unnecessary and distracting footnotes.

There may be a place for footnotes in a scholarly monograph, but in a popular book they are usually signs of sloppy writing. If they say something important they should be incorporated into the text, otherwise the casual reader may miss something vital. If they are not essential, they should be left out for fear of muddying the water.

An example from this book relates to Arthur Eddington's eclipse expedition of 1919 to the West African island of Principe, where he made the first measurement of the deflection of light by the Sun, predicted by Einstein's general theory of relativity. Afterwards, Eddington wrote a poem containing the phrase "light has weight". In a footnote, Seife claims that this is misleading because light "does not actually have mass". In everyday language, mass and weight are more-or-less synonymous, but any high-school physics student knows that these terms have quite different meanings in the language of classical mechanics. Eddington knew the difference too. In newtonian language, weight is a measure of the gravitational force on a body. A massive body can be weightless, if it is in freefall or in a region without a gravitational field. On the other hand, in Einstein's theory, massless particles such as photons can feel the effects of gravity, so it is reasonable to describe them as having weight. No poetic licence is required, and I'm not sure Eddington possessed one anyway.

This may seem a pedantic objection, but my grumble is less about the fine distinction between the concepts of weight and mass as about the pointlessness of raising the issue in the first place. Besides, errors of fact are even less forgivable than errors of judgement: the famous eclipse mentioned above happened on 29 May 1919, not 26 March, as stated by Seife.

I can offer a useful general tip about popular-science books: stop reading immediately when you come across the word 'mind-boggling'. This is the point where the author admits defeat, so it's only fair for the reader to do likewise. Applying that principle to this book will get you about half-way or, on a scale from alpha to omega, about as far as mu. ■

Peter Coles is in the School of Physics and Astronomy, University of Nottingham, University Park, Nottingham NG9 2HL, UK. He is the author of *Cosmology: A Very Short Introduction*.

Science in culture

Pixels and piety

The digital collection at the Museum of the History of Science in Florence.

Martin Kemp

Many museums have spent large sums of money embracing the digital age, often to no great effect. Online access to images and information is certainly valuable in extending the audience of any museum, but most of the projects go no further than archiving, and make little creative use of the potential of digital imaging. Likewise, most of the on-site digital access that museums provide for their visitors relies on low-level interactivity conceived by middle-aged curators who hope that touching a computer screen will transform the museum experience for 'young people'.

Happily, a few museums are now moving on to a more creative level to enhance their interaction with both real and virtual visitors. No museum has a more ambitious programme in this respect than the Museum of the History of Science in Florence, Italy, whose director, Paolo Galluzzi, is developing an imaginative range of digital access (as the site map at galileo.imss.firenze.it testifies). The most ambitious of the projects, galileo/thek@, will, when it is completed in February 2004, provide the most comprehensive set of images, primary sources, interpretative materials and animations for any historical figure. Animations, of which there are several on the museum's website, are particularly important, as they provide an immediate and dynamic insight into the working of instruments that sit as inertly as sports trophies in their sealed museum cabinets.

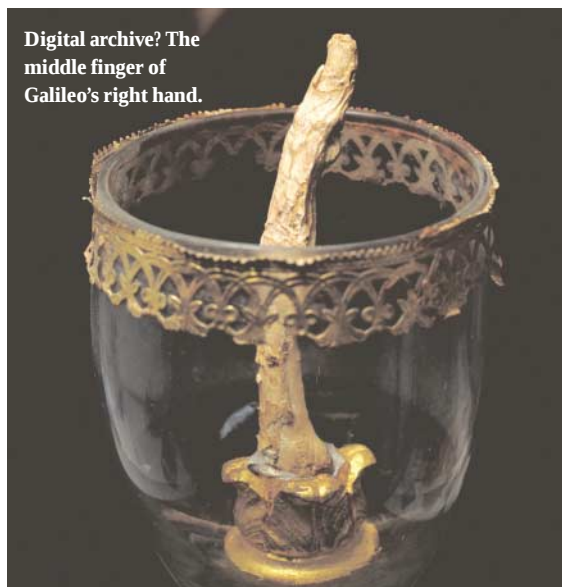
In Florence, the latest technologies are blended with extraordinary levels of traditional reverence for Galileo, the 'god' of Tuscan science. Alongside hard-nosed information about galilean astronomy and dynamics, scholarly information about manuscripts and editions, and extensive secondary literature, are some extraordinary memorabilia of the Italian scientist, presented with a piety that is at least the equal of any found in a church. The most sanctified relic is the shrivelled middle digit of Galileo's right hand — the medium of which is laconically listed in the museum catalogue as "finger".

When Galileo's remains were transferred to the main body of the Florentine church of Santa Croce on 12 March 1737, the antiquarian, Anton Francesco Gori, took the opportunity to detach the finger as if it were a revered fragment from the corpse of a saint. For many years, the relic was exhibited in the Biblioteca Laurenziana, having acquired its elaborate eighteenth-century mount

and inscription, before passing in 1841 to the new Tribuna di Galileo in the Museo di Fisica e Storia Naturale on the via Romana in Florence, and eventually to its current resting place.

The flavour of such piety is embodied in another Galileo reliquary now housed in the same museum. This contains the objective lens used by the astronomer in 1610 to discover the moons of Jupiter, which he designated the 'Medicean Planets'. Mounted in a florid ivory frame by Vittorio Croster in 1677, it was for years part of the cherished collections of the Medicean Grand Dukes of Tuscany in the Galleria degli Uffizi, alongside the masterpieces of Leonardo, Michelangelo and Raphael.

Digital archive? The middle finger of Galileo's right hand.



One of the Latin inscriptions on the mount of the lens captures perfectly the tone of adulation: "The sky, opened by the lynx-like mind of Galileo with this first lens of glass, showed stars never before seen, rightly called Medicean by their discoverer. The knowing mind masters even the stars, indeed." The animal allusion is to the Accademia Nazionale dei Lincei, the pioneering scientific academy of the 'lynx-eyed', of which Galileo was the brightest luminary. The academy celebrates its 400th anniversary this year (*Nature* **422**, 467–468, 2003).

The finger may be dry and withered, and the lens cracked beyond even rudimentary utility, but the relics maintain their potency. They are, as they say, the real thing. At the end of the day, this possession of the authentic item provides the enduring strength and fascination of museums. Digitization is a potentially vivid means of enhancing the relationship between the museum object and the curious visitor. But it is not an end in itself.

Martin Kemp is professor of the history of art at the University of Oxford and co-director of Wallace Kemp/Artakt.

How animals work

Summarize yourself in the form of a title of a paper in Nature.

Cool curiosity surviving despite frigid environment.

What was your first experiment as a child?

Collecting all kinds of insects, in particular chrysalises to watch butterflies blossom out, and beetles. One evening my father brought home a hornet he had captured in his coat. The hornet escaped, causing panic. As a result, my mother didn't like entering my room, which was quite a zoo.

Who has been the most important mentor in your career?

My first mentor was Jean Dorst, author of *Avant que Nature Meure*. When I was 15, I wrote to him about my dream to become an ornithologist. His reply emphasized the difficulty of making a living at it, but he encouraged my passion for fieldwork, while understanding my dissatisfaction with an academic science that I wished would be less descriptive.

Whose graduate student would you most like to have been?

Charles Darwin, when he was working in the Galapagos, to help him label his samples properly.

What single scientific paper or talk changed your career path?

'Animal anorexias' (N. Mrosovsky and D. F. Sherry *Science* **207**, 837–842; 1980) showed that top science can be achieved with good questions and very simple measurements. This paper was also stimulating at a difficult time, when a scientific director of the CNRS told the community that 'physiology' had become meaningless as a field.

What book has been most influential in your scientific career?

Reading *How Animals Work* by Knut Schmidt-Nielsen (Cambridge University Press, 1972), I understood I had found the direction I was searching for. Not long after, I had the privilege of working in his lab. A master and my mentor in the understanding of function in nature, he also taught me how to say what is interesting in words that are accessible to a large audience.

What gives you the most job satisfaction now? What are your major frustrations?

The most satisfaction comes at the time an interesting question is unravelled, and also when students or postdocs have their work accepted for publication or obtain a good position. Frustration comes with the continual increase in both scientific bureaucracy and the complexity of accounting.

What book is currently on your bedside table?

La Tapisserie de Bayeux (Editions du Zodiaque, 2002) by Lucien Musset from the University of Caen. He presents a new approach to this 65-metre tapestry that illustrates the turbulent accession of William Duke of Normandy to the throne of England in the years 1064–66.

What music heads the playlist in your car or lab?

Mozart, Vivaldi and world music, particularly from the Middle East, China and the Andes.

Assuming the dead can be raised and/or time travel exists, with whom from the world outside science would you most like to have dinner?

Jean-François Champollion, who deciphered hieroglyphics using the Rosetta Stone. Although I studied Latin and ancient Greek, I am fascinated by the many ancient languages Champollion knew when he was very young.

Where and when would you most like to have lived or worked?

Now is fine. But I should like to return in several hundred years to see how humans have handled natural resources.

What do you most dislike about having research published?

When a paper appears at the same time as one from a competitor. To be simultaneous is to be too late.

What's the best piece of advice you've received?

Be yourself. Don't always follow the trends.

What do you do to relax? (Remember, Nature is a family journal.)

Run for half an hour every morning at sunrise, on a path among ponds with lots of diving and dabbling birds, listening to the tapping noise of woodpeckers.

What would you have become, if not a scientist?

A war reporter emphasizing human suffering, or a doctor in the wilderness.

What overlooked or underrated discovery really changed the science in which you work?

Long-term data collection on wildlife, which today is extremely useful for studying the impact of climate change.

Do you have a burning ambition to do or learn something of no practical or immediate value? If so, what?

Like Champollion, to become fluent in ancient languages.

What do you want written on your gravestone?

'Ideas still available for the next three decades.'



Yvon Le Maho

Yvon Le Maho studies how animals adapt to the environment, and is especially fond of penguins. He is director of the CNRS Centre d'Ecologie et Physiologie Énergétiques in Strasbourg, France.

Under what conditions do you have your greatest and most inspired ideas?

On a mountaintop, or contemplating icebergs trapped by winter sea ice.

Is there a 'tyranny of reductionism' in how scientists are trained today?

Reductionist tyranny is not in training, but hidden behind statements such as 'from genes to function' with the agenda that they represent the only route towards integrative biology.

What's the one thing about science that you wish the public understood better?

That scientists aren't always able to give clear-cut answers.

What music would you have played at your funeral?

Gregorian chant, because of its peaceful beauty, and some cantigas of the Medieval Mediterranean, as these religious songs express a joy of life stronger than death.

If you were reborn as a comic-book superhero, what would be your superhuman power?

Flying with migrating birds.

You're the only scientist at a party. How do you describe what you do for a living?

I don't describe but I ask simple questions, such as: 'Do you think that there will still be wild fish in the sea next year?'

You've just been told (in confidence) that the world will end tomorrow. What do you do next?

You said *Nature* was a family journal, didn't you...? ■

Spin control for asteroids

Richard P. Binzel

Random collisions between asteroids would seem to cause their spin axes to be tilted in all directions. Surprisingly, the gentle recoil force of thermal re-radiation may bring their spin axes into alignment.

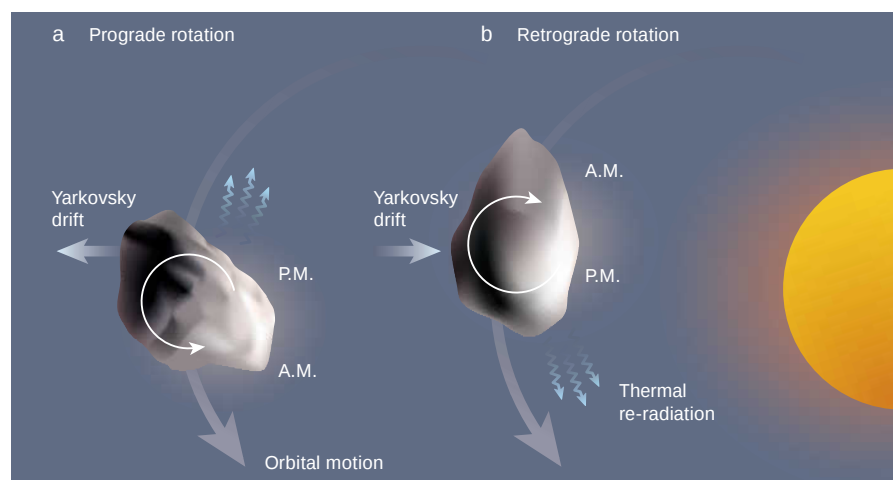


Figure 1 The Yarkovsky effect. An asteroid is warmed by sunlight, its afternoon side becoming hottest. As a result, that face of the asteroid re-radiates most thermal radiation, creating a recoil force on the asteroid and causing it to drift a little. The direction of the radiation depends on whether the asteroid is rotating in a prograde (anticlockwise) manner (a) or in a retrograde (clockwise) manner (b).

Influencing the public perception of a political story might seem like small potatoes compared with controlling the spin of the mountain-sized rocks in space that we call asteroids. These leftover building blocks from the era of planetary formation have undergone random and relentless collisions that have sculpted their shapes over the past 4.5 billion years. These bumps in the night should result in random orientations of the asteroids' spin axes, yet observations have revealed unexpected spin alignments¹. On page 147 of this issue, Vokrouhlický *et al.*² show how the slow but steady recoil force of thermal re-radiation can win out over the heavy hand of collisions in controlling the direction of asteroid spins.

Planetary bodies such as asteroids maintain an equilibrium temperature by re-radiating (at thermal infrared wavelengths) the same amount of energy that they absorb from sunlight. In about 1900, I. O. Yarkovsky^{3,4}, a Russian engineer, wrote about the effect that this recoil could have on the motion of a planetary body (Fig. 1). The 'Yarkovsky effect' relies on the same principle that drives more swimmers to the beach in the afternoon than in the morning — the afternoon side of a rotating planet is hotter as a consequence of having absorbed a full day's worth of sunshine. Being hottest, the afternoon side therefore produces the greatest amount of thermal re-radiation. Yarkovsky

reasoned that the recoil from this one-sided thermal re-radiation could preferentially slow down or speed up the orbital velocity of an asteroid, depending on whether the tilt of its spin axis meant that its afternoon side faced forwards or backwards. The resulting 'Yarkovsky drift' is most effective for small asteroids and might be a key component in the delivery of meteorite samples from the asteroid belt to the Earth⁵.

Thermal re-radiation was also recognized to have some effect on the spin rate and orientation of an asteroid, dubbed the 'YORP effect' after the names of researchers who detailed the mechanics involved⁶ (Yarkovsky, O'Keefe, Radzievskii and Paddick). The YORP effect, acting over millions of years, might have asteroids dancing the 'hokey-cokey' (also known as the 'hokey-pokey') — spinning them up, spinning them down and turning their spin axes around — but there was no reason to think that any particular rotation state would be preferred. Besides, random collisions would shake them all about and start the dance all over again.

Theory often drives observations, but this time observations led the way. Groups of asteroids called 'families' have similar orbits resulting from the collisional disruption of a larger parent body. An early study⁷ of asteroids in the Koronis family, which are nearly three times as far as Earth from the Sun, revealed unusual rotation characteristics. Measurements revealing how Koronis family

asteroids reflect sunlight over the course of a rotational cycle (called a 'light curve') showed a greater average variation in brightness than for asteroids in other families or similar-sized non-family asteroids. Two hypotheses were offered⁷: either the formation of the Koronis family resulted in more elongated shapes (and therefore greater variability in brightness) or the spin axes of Koronis family members were preferentially aligned nearly perpendicular to their orbit plane — an orientation that would always display a nearly equatorial view of the asteroid (and hence maximum variability of brightness) to Earth-based observers.

Although early follow-up studies⁸ were consistent with a preferential alignment, a decisive test of this hypothesis was a gargantuan task. From Earth, an asteroid's spin vector can be divined only by measuring its light curve from many different viewing angles. For a Koronis family asteroid, an observation opportunity comes only once every 15 months, and five or more sets of observations, all at different angles, might be

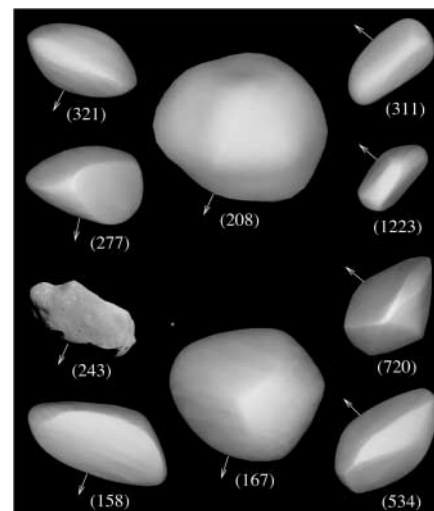


Figure 2 Spin up, spin down. The spin axes of the Koronis family of asteroids are very closely aligned^{1,10}, pointing in one of two directions depending on whether the asteroid spins clockwise (downward arrows) or anticlockwise (upward arrows). Vokrouhlický *et al.*² show that these 20–40-km-diameter asteroids could have been nudged into these aligned states by a featherweight recoil force created by thermal re-radiation from the hot afternoon side of each asteroid, acting over a period of 2–3 billion years. (Reproduced with permission from ref. 10.)

required. At each viewing opportunity, many dozens of meticulous measurements are needed over the typical eight-hour rotation period of the asteroid to define its light curve; light curves must be measured repeatedly over several weeks to determine the asteroid's spin period precisely. Now multiply this effort for one asteroid by at least ten, to get a reasonable statistical sample, and... you get the idea.

But someone was up to the challenge. Using the Wallace Astrophysical Observatory's 0.6-m telescope for nearly a decade, postgraduate student Stephen Slivan and a cadre of undergraduate students amassed thousands of measurements for ten Koronis family asteroids^{9,10}. Decidedly nonrandom spin vectors were found¹. Four prograde (anticlockwise) rotators have nearly aligned rotation axes with very similar spin periods of about eight hours, and the remaining six retrograde (clockwise) rotators are also nearly co-aligned together (but on a different heading) with a dispersion of spin periods between 3 and 30 hours (Fig. 2). The alignments almost defied belief — how could a collisionally dominated family have aligned spin vectors? But this *tour de force* data set¹ (confirmed by an independent analysis of spin vectors by Mikko Kaasalainen^{10,11}) called out for explanation.

Vokrouhlický *et al.*² have responded to the call, with a model that invokes the YORP effect and the influence of Solar System planets. For the 20–40-km-diameter members of the Koronis asteroid family (the same size range and location as the Slivan sample), they find that prograde rotators should have their spins slowed and their spin axes cast into a slow precession like a top. When the precession rate matches the rate at which the plane of Saturn's orbit undergoes a slow twist, the asteroid becomes caught in an equilibrium state that Vokrouhlický *et al.* call a 'Slivan state'. The spin vector and spin rate become fixed, and their calculated values closely match those for the observed sample. Retrograde rotators don't seem to be influenced by Saturn's twist or to land in an equilibrium state, but thermal torques (turning forces) turn their spin vectors upside down with rotation rates diverging towards either fast or slow values — again just as is observed. According to Vokrouhlický and colleagues' model, it takes typically 2–3 billion years for the Koronis family asteroids to evolve to their end states, with smaller and more irregularly shaped objects proceeding more quickly.

How well this YORP-based model² holds up can be tested through further observations. Most asteroids smaller than 40 km in diameter should be influenced by the effect, with those in the outer asteroid belt being most likely to land in observable Slivan states. Measuring an abundance of upside-down spin vectors for fast or slowly rotating asteroids would be another tell-tale

fingerprint of the YORP effect. If it can be shown to dominate asteroid spins, the implication is that geologically frequent collisions are highly inefficient in transferring rotational angular momentum. Collisions that fracture but do not destroy asteroids might leave them so weak that a hard push has little effect. In contrast to what has been long thought, a light touch might be the best way to influence asteroid spins. ■

Richard P. Binzel is in the Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.
e-mail: rpb@mit.edu

- Slivan, S. M. *Nature* **419**, 49–51 (2002).
- Vokrouhlický, D., Nesvorný, D. & Bottke, W. F. *Nature* **425**, 147–151 (2003).
- Ópik, E. J. *Proc. R. Irish Acad.* **54**, 165–199 (1951).
- Hartmann, W. J. *et al. Meteoritics Planet. Sci.* **34**, 161–167 (1999).
- Bottke, W. F., Vokrouhlický, D., Rubincam, D. P. & Broz, M. in *Asteroids III* (eds Bottke, W. F. *et al.*) 395–408 (Univ. Arizona Press, Tucson, 2002).
- Rubincam, D. P. *Icarus* **148**, 2–11 (2000).
- Binzel, R. P. *Icarus* **73**, 303–313 (1988).
- Binzel, R. P. in *Asteroids, Comets, Meteors III* (eds Lagerkvist, C. I. *et al.*) 15–18 (Uppsala Universitet, 1990).
- Slivan, S. M. & Binzel, R. P. *Icarus* **124**, 452–470 (1996).
- Slivan, S. M. *et al. Icarus* **162**, 285–307 (2003).
- Kaasalainen, M., Mottola, S. & Fulchignoni, M. in *Asteroids III* (eds Bottke, W. F. *et al.*) 139–150 (Univ. Arizona Press, Tucson, 2002).

Ageing

A toast to long life

Toren Finkel

Reducing food intake increases lifespan in many species. A small molecule that occurs naturally in plants seems to mimic the beneficial effects of caloric restriction and extend longevity in yeast.



Figure 1 **A quest for longevity.** Five hundred years ago, the Spanish explorer Ponce de León drank his way around the Florida coast during his expedition to find the legendary fountain of youth.

In the spring of 1512, three wooden vessels commanded by the Spanish explorer Ponce de León left the warm Caribbean waters off Puerto Rico in search of the fountain of youth. Fuelled by his desire for immortality, Ponce de León was convinced that drinking from this legendary spring would confer a state of eternal youth. Local legend suggested that the fountain could be found in the lands to the north and that it was surrounded by magnificent flowering plants. Over the next few months, the explorers travelled from island to island, tasting rivers and lakes (Fig. 1) until short supplies and hostile natives forced them to abandon their quest.

Today, the once desolate shores of Florida

— the 'land to the north' that Ponce de León discovered inadvertently on his voyage — are crowded with ageing retirees who yearn just as passionately for the magical elixir of youth. Although the past two decades have seen an explosion in our understanding of the molecular regulation of ageing, a simple potion or chemical that could slow the ageing process seems as elusive today as it was 500 years ago. But on page 191 in this issue, Howitz *et al.*¹ present a study of longevity in yeast that could represent the first step towards creating such a hypothetical elixir.

It is well established that reducing food intake (caloric restriction) extends lifespan in a wide range of species². But given the current epidemic of obesity in industrialized

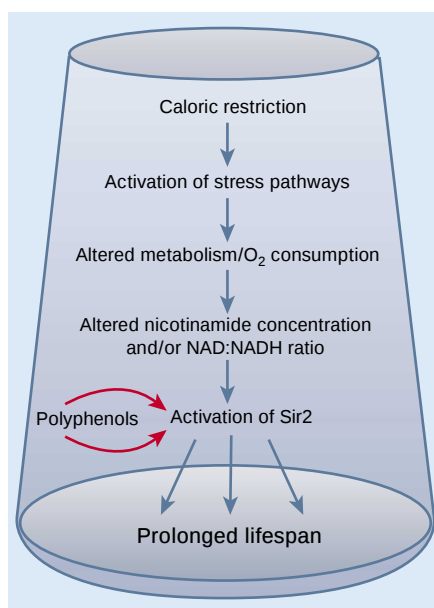


Figure 2 The pathway to long life. When yeast cells are deprived of food (caloric restriction), stress pathways are activated and the cells are forced to derive energy from alternative substrates. This produces alterations in oxygen consumption, which in turn affects the ratio of oxidized to reduced forms of nicotinamide adenine dinucleotide (NAD:NADH) or the concentration of its derivative nicotinamide. NAD stimulates the activity of Sir2, which in turn chemically modifies several proteins that are involved in cellular processes affecting longevity. Howitz *et al.*¹ have found that plant polyphenols directly activate Sir2 and seem to mimic the beneficial effects of food restriction. Related pathways may exist in higher organisms.

societies, eating less for the sake of longevity is not likely to gain widespread compliance. More palatable would be a drug that could mimic the effects of caloric restriction — a chemical that would allow the individual to eat normally, while tricking the body to respond as though food were in short supply. Such mimetics might be thought of as the pharmaceutical equivalent of ‘eating your cake *without* having it’.

Studies of longevity in simple organisms have greatly expanded our understanding of how caloric restriction might increase lifespan. In the budding yeast *Saccharomyces cerevisiae*, nutrient withdrawal extends longevity through a pathway that requires the enzyme Sir2 (ref. 3). Overproducing this enzyme can prolong the life of yeast grown under normal nutrient conditions⁴. Similarly, in the evolutionarily more advanced worm *Caenorhabditis elegans*, increased expression of the worm’s version of Sir2 has also been shown to extend lifespan⁵.

The Sir2 enzyme belongs to a large family of evolutionarily conserved molecules termed sirtuins. In lower organisms, such as yeast and worms, these enzymes regulate a

wide range of cellular activities that affect lifespan, including modulating how tightly DNA is packaged inside cells. In mammalian cells, sirtuins act as regulators of programmed cell death and differentiation (cell maturation)⁶. Sirtuins exert their effects on these cellular processes by removing acetyl groups from specific target proteins. Interestingly, this ‘deacetylase’ function depends on the intracellular concentration of a molecule involved in metabolism — nicotinamide adenine dinucleotide (NAD). This molecule can exist in two states, oxidized and reduced, and it is the oxidized form that greatly enhances Sir2 activity. It seems that in yeast, caloric restriction may regulate Sir2 activity, and hence prolong life, by subtly shifting the ratio of oxidized to reduced NAD or by altering the level of the NAD derivative nicotinamide^{7,8}. Together, these findings suggest a potential mechanism by which metabolic activity and lifespan might converge (Fig. 2).

Building on the knowledge that caloric restriction prolongs longevity through Sir2, Howitz *et al.*¹ searched for a small molecule that could activate this enzyme directly. Using several chemical ‘libraries’, these investigators discovered two related compounds that each stimulated Sir2 activity. Both compounds belong to a family of molecules called polyphenols — products of metabolism in plants. One of the most widely studied of these compounds is resveratrol, a plant polyphenol that is abundant in red wine and is reputed to underlie many of wine’s health-related benefits. Interestingly, resveratrol seemed to be the most potent Sir2 activator of all of the plant polyphenols tested. The authors showed that this chemical prolonged the lifespan of yeast by approximately 70%. The extension of longevity was entirely dependent on Sir2 — yeast strains deficient in this enzyme did not benefit from resveratrol treatment.

Could plant polyphenols be the long-

sought elixir of youth? Previous studies have hinted that these compounds have several potential health benefits, especially in protecting against age-related maladies such as cancer, neurodegeneration and atherosclerosis⁹. Interestingly, caloric restriction is also thought to protect against these diseases. But caution is warranted before endorsing a strict Cabernet Sauvignon-based regimen. First, the concentration-dependent effects of resveratrol as observed by Howitz *et al.* were complicated. At relatively low doses these molecules stimulated sirtuin activity, but, at least in certain assays, higher doses had the opposite effect. This is not an ideal characteristic for a pharmaceutical drug. Second, and more importantly, life extension in yeast is a long way from life extension in higher organisms. Indeed, how sirtuins function in mammalian ageing is not yet known.

Further unravelling of the molecular signalling pathways that accompany caloric restriction should provide clues to other potential targets for drug development. In a strange way, however, the study by Howitz *et al.* suggests that Ponce de León’s misbegotten quest for a fountain of youth surrounded by flowering plants was not so delusional after all. The explorer’s only mistake was that he kept sampling the waters, when he should have been testing the plants. ■

Toren Finkel is at the Cardiovascular Branch, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, Maryland 20892-1622, USA.
e-mail: finkelt@nih.gov

1. Howitz, K. T. *et al.* *Nature* **425**, 191–196 (2003).
2. Masoro, E. J. *Exp. Gerontol.* **35**, 299–305 (2000).
3. Lin, S.-J., Defossez, P. A. & Guarente, L. *Science* **289**, 2126–2128 (2000).
4. Kaerberlein, M., McVey, M. & Guarente, L. *Genes Dev.* **13**, 2570–2580 (1999).
5. Tissenbaum, H. A. & Guarente, L. *Nature* **410**, 227–230 (2001).
6. Denu, J. M. *Trends Biochem. Sci.* **28**, 41–48 (2003).
7. Anderson, R. M. *et al.* *Nature* **423**, 181–185 (2003).
8. Lin, S.-J. *et al.* *Nature* **418**, 344–348 (2002).
9. Tapiero, H., Tew, K. D. & Matthe, G. *Biomed. Pharmacother.* **56**, 200–207 (2002).

Condensed-matter physics

Vortices and hearts

John Clarke

A single vortex of flux, formed inside a superconducting Josephson junction, has been detected undergoing quantum tunnelling — a feature that could be developed into a quantum bit.

Vortices are ubiquitous — from the mythical whirlpool of Charybdis in the Strait of Messina, to the putative cosmic strings that were frozen into spacetime shortly after the Big Bang. In superconductors, too, swirling currents generate vortices of flux. On page 155 of this issue, Wallraff *et al.*¹ show that a single flux vortex can undergo macroscopic quantum

tunnelling² to escape from a potential well inside a long Josephson junction; they also show that the vortex’s energy in the controllable well is quantized. These observations of uniquely quantum phenomena mean that this system becomes another entrant on the growing slate of candidates to make superconducting ‘qubits’ for quantum computing.

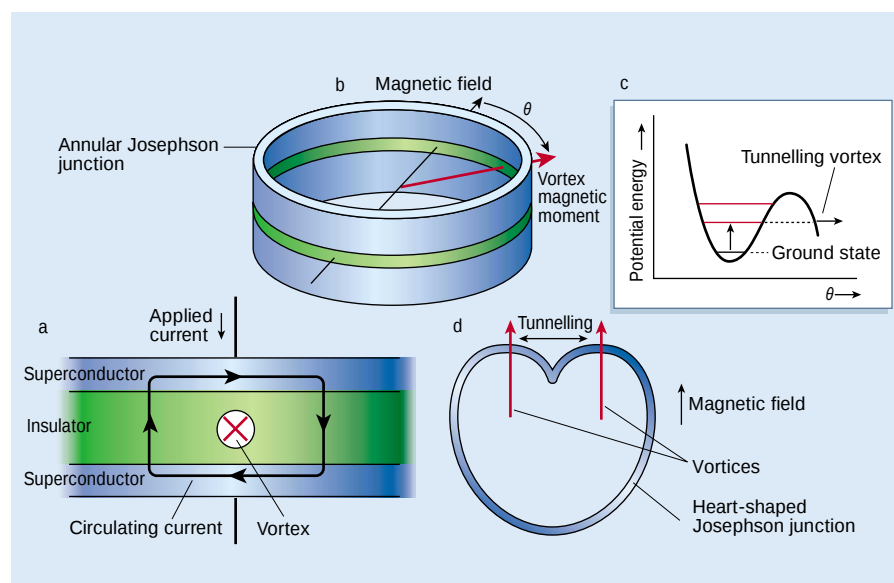


Figure 1 Quantum tunnelling and the single vortex. a, The long Josephson junction (shown here in cross-section) consists of two superconducting films separated by an insulating layer through which Cooper pairs of electrons can tunnel. The circulating supercurrent creates a vortex. b, In the annular version of the junction¹, the magnetic moment of the vortex is at an angle θ to the direction of an applied magnetic field. c, As θ changes, the potential energy of the vortex varies periodically (with $\cos \theta$), reaching a minimum when the magnetic field and moment are aligned. A current applied across the junction tilts the potential well, and microwave radiation excites the vortex from its ground state to a higher energy level from which it is more likely to escape the potential well through tunnelling. d, If the annular Josephson junction were instead a heart-shaped junction, with two minima of potential energy, the system could make a useful quantum bit: a single vortex could tunnel between states in the two wells, at any instant being in a superposition of both states.

Two kinds of vortex can develop in a superconductor. The first, the Abrikosov vortex, penetrates certain (type II) superconductors above a critical value of applied magnetic field. It can hop between pinning sites under thermal activation, causing dissipation in current-carrying wires and generating noise in sensors such as superconducting quantum interference devices (SQUIDs). Quantum tunnelling of Abrikosov vortices remains controversial. Wallraff and colleagues' work¹, however, is concerned with the second kind of vortex, the Josephson vortex. This exists in a Josephson junction, formed by a sandwich of two superconducting layers and a thin insulating layer, through which electrons in the form of Cooper pairs can tunnel coherently (Fig. 1a).

In an elegant twist on the conventional linear junction, Wallraff *et al.* made an annular junction between two narrow rings of the superconductor niobium, stacked one on top of the other (Fig. 1b). The flux in each ring is quantized in units of $h/2e$ (where h is Planck's constant and e is the charge on an electron); when this flux quantum number differs between the two rings by one unit, the difference is manifested as a single vortex in the junction. If a magnetic field is applied in the plane of the annulus, at an angle θ to the magnetic moment of the vortex, the potential energy of the system is proportional to the cosine of θ (Fig. 1b) — so the energy of the vortex is periodic in θ , with a minimum

when the applied field and the vortex moment are parallel.

If an external current is then applied to the junction, across the two superconducting layers, the resulting imbalance in the tunnelling currents produces a force on the vortex, tilting its potential well and lowering — but not removing — the barrier to escape. At high temperature, in the classical regime, the vortex can escape from this well, lifted by thermal energy over the barrier it forms. At low temperature in the quantum regime, however, the vortex escapes by macroscopic quantum tunnelling through the barrier (Fig. 1c).

Wallraff *et al.*¹ investigate the escape process by ramping up the current that is flowing through the annular junction until they see a jump in the voltage across the junction, which corresponds to the vortex beginning to rotate rapidly around the annulus (the voltage is proportional to the vortex's velocity). This 'depinning' of the vortex, and its subsequent escape from the well, is a stochastic process, so repeated measurements yield a distribution of switching currents at which depinning occurs. As the temperature is lowered, the width of this distribution shrinks, indicating a crossover from the classical regime to macroscopic quantum tunnelling in the quantum regime.

A second demonstration of the quantum nature of the vortex is the quantization of its energy in the potential well³. The applied

magnetic field and current can be adjusted so that the well contains (say) three energy levels; absorbing radiation of specific frequencies will induce the vortex to make a transition between them. Typically, the transition frequency from the ground (lowest) state to the next energy level up is about 10 GHz, so that at a temperature of, say, 25 mK — corresponding to a frequency of about 0.5 GHz — the vortex is virtually certain to be in its ground state. The probability that the vortex will tunnel from the ground state is low, because the barrier is relatively high and wide (Fig. 1c). But irradiating the system with 10-GHz microwaves causes the vortex to make a transition to its excited state, from which the tunnelling escape rate is much higher. By observing a peak in the escape rate as a function of microwave frequency, Wallraff *et al.* demonstrate that the energy of the vortex in the well is quantized, and depends on the magnetic field and current exactly as predicted.

These observations of macroscopic quantum tunnelling and quantized energy levels in a Josephson junction are in themselves not new. In fact, they were first demonstrated in a small, current-biased junction³, and then in an 'rf SQUID' — a Josephson junction shunted by a superconducting loop⁴. But the experiment performed by Wallraff *et al.*¹ is unique in that it clearly addresses the behaviour of a single vortex. Consequently, an exciting direction to explore with this annular junction is its potential as a qubit in a quantum computer.

There have been rapid developments in superconducting qubits, in which two states of a variable must exist in a quantum superposition. This superposition has been demonstrated in a single, current-biased Josephson junction, where it involves two neighbouring energy levels in the well⁵; in a SQUID, where it involves the up and down flux states^{6,7}; and in the charge of a tiny superconducting island, where it involves two states differing by a single Cooper pair^{8,9}. Evidence for the entanglement of two superconducting qubits has also been reported^{10,11}.

How would one make a Josephson-vortex qubit? One possibility is to base it on two of its quantized energy levels, analogous to the single-junction qubit⁵. A more intriguing idea, already suggested by Wallraff *et al.*¹², is to make a heart-shaped qubit (Fig. 1d). The two lobes, combined with the applied field, create a double-well potential between which the vortex tunnels, in analogy with the SQUID qubit. Superposition of the two states resolves their degeneracy, forming two distinct energy levels. To evaluate the usefulness of the vortex qubit, its relaxation time and decoherence time should be measured — that is, how long it remains in the excited state and how long the coherent quantum superposition lasts. As with other superconducting qubits, the enemies are

noise, dissipation and resonances in the environment and in the structure itself. Beating these down remains a great challenge for the field.

John Clarke is in the Department of Physics, University of California, Berkeley, and the Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA.

e-mail: jclarke@physics.berkeley.edu

1. Wallraff, A. *et al.* *Nature* **425**, 155–158 (2003).
2. Caldeira, A. O. & Leggett, A. J. *Ann. Phys.* **149**, 374–456 (1983).
3. Martinis, J. M., Devoret, M. H. & Clarke, J.

Phys. Rev. Lett. **55**, 1543–1546 (1985).

4. Rouse, R., Han, S. & Lukens, J. E. *Phys. Rev. Lett.* **75**, 1614–1617 (1995).
5. Martinis, J. M., Nam, S., Aumentado, J. & Urbina, C. *Phys. Rev. Lett.* **89**, 117901 (2002).
6. Friedman, J. R., Patel, V., Chen, W., Tolpygo, S. K. & Lukens, J. E. *Nature* **406**, 43–46 (2000).
7. Chiorescu, I., Nakamura, Y., Harmans, C. J. P. M. & Mooij, J. E. *Science* **299**, 1869–1871 (2003).
8. Nakamura, Y., Pashkin, Yu. A. & Tsai, J. S. *Nature* **398**, 786–789 (1999).
9. Vion, D. *et al.* *Science* **296**, 886–889 (2002).
10. Pashkin, Yu. A. *et al.* *Nature* **421**, 823–826 (2003).
11. Berkley, A. J. *et al.* *Science* **300**, 1548–1550 (2003).
12. Wallraff, A., Koval, Yu., Levitchev, M., Fistul, M. V. & Ustinov, A. V. *J. Low Temp. Phys.* **118**, 543–553 (2000).

Behavioural ecology

Father knows best

Paul W. Sherman and Bryan D. Neff

When females mate with several males, paternity may be uncertain. But male savannah baboons are seldom confused: when intervening in fights between youngsters, they generally support their own offspring.

Behavioural ecology is the study of how natural selection shapes behaviour in relation to ecological and social conditions. A fundamental principle is that individuals should promote the spread of their own genes over those of competitors. For example, adults sacrifice themselves to defend and nourish offspring. But they try to avoid providing assistance to non-relatives (that is, they avoid ‘misdirecting’ their nepotism), usually by guarding their mate against other sexual liaisons and attempting to distinguish their own offspring from unrelated young¹.

Mistakes in discrimination are most likely for males of internally fertilizing species, especially when the females accept multiple partners. Indeed, in most such species males do not care for young, and it has been suggested that this is because they are uncertain which juveniles are their own. But a new study by Buchan and colleagues², described on page 179 of this issue, reveals that when male savannah baboons (*Papio cynocephalus*) intervene in fights between juveniles, they favour offspring over non-kin. This work is important because it reveals that such intervention is a manifestation of nepotism, that paternal care can occur even when females copulate with several different males, and that the kin-recognition mechanisms of a wild primate approach those of humans in precision and sophistication.

In nature, multiple mating is commonplace and widespread, and molecular genetic analyses reveal that it typically results in females bearing the offspring of several males^{3,4}. To avoid misdirecting nepotism, males generally rely on indirect cues of their paternity of entire broods⁵. For example, male bluegill sunfish (*Lepomis macrochirus*) are less willing to defend eggs and fry against

predators, and more likely to abandon or cannibalize young, if cuckolders had lurked nearby during spawning (a proven threat to paternity) than if they had experienced no cuckolders⁶. However, there has been no evidence that bluegill sunfish, or males in any other wild vertebrate, can distinguish offspring from unrelated juveniles within a given brood⁷—until now.

Buchan and colleagues’ study² builds upon 30 years of continuous research on the baboons of Kenya’s Amboseli Basin, located at the foot of Mount Kilimanjaro. These primates live in multi-male, multi-female troops, in which adult males frequently intervene in fights that break out between juveniles. Such intervention and physical support shields young from injury and stress, and enhances their dominance rank. During 1999–2002, the researchers recorded which youngster was assisted when an adult

male baboon intervened in a fight. By analysing DNA from faecal and blood samples the authors established, *post hoc*, whether there was a genetic relationship between each male–juvenile dyad. They report that when males were faced with a choice, they were significantly more likely to support offspring than unrelated juveniles (Fig. 1). Male support of females that are being attacked is a well-known indicator of ‘friendship’; it now seems that such behaviour toward youngsters is in fact paternal care.

But how do male baboons avoid misdirecting their nepotism? Two mechanisms are possible. First, males might rely on indirect cues of paternity, as with the male bluegill sunfish that notice cuckolders during spawning. Female baboons are most fertile during the last five days of the follicular phase of their oestrous cycle, which is recognizable by the degree of swelling of their sexual skin. Males might therefore behave according to a simple darwinian algorithm (behavioural rule), such as ‘assume that a juvenile is your offspring if you frequently copulated and consorted with (guarded) its mother when her perineum was maximally swollen’. If so, males should distinguish such ‘behaviourally predicted’ offspring from unrelated juveniles born to females with whom they did not copulate. Moreover, the probability of a male assisting a juvenile should correlate with the proportion of time that the male spent sexually monopolizing its mother during her period of maximum fertility. This is indeed what Buchan *et al.* observed.

A second, complementary recognition mechanism involves a male using direct (phenotypic) cues, such as how a juvenile looks or smells, to assess his likelihood of being its father. This mechanism, known as phenotype matching, is widespread in nature⁸, including among primates. In humans, the resemblance of a baby to its mother’s husband is a well-known source



Figure 1 **Paternal love:** a male baboon intervenes in a squabble between two juveniles on behalf of his own offspring.

of interest and familial controversy. But to demonstrate convincingly that males in a given species use self-referential phenotype matching requires researchers to identify the phenotypic cue, alter or obliterate it, and then observe predictable changes in — and even failure of — recognition. Unfortunately, such manipulations are not feasible for wild baboons, and they might also disrupt the 'normal' behaviour of the animals.

Nonetheless, Buchan and colleagues' data indicate that male baboons do use direct as well as indirect cues. As mentioned above, when males intervened in fights between the offspring of females with whom they had and had not consorted, they (the males) consistently favoured the offspring of the former females. In this case, both types of cue — indirect and direct — would be in agreement that one youngster is an offspring and the other is not. But when males intervened in fights between offspring of different consorts, and only one juvenile was actually sired by the male, direct and indirect cues must have been in conflict. Yet, although the males' choices were less decisive, they still favoured true (genetic) offspring over unrelated juveniles. This implies that males do use phenotype matching, and that when direct and indirect cues conflict, the direct cues 'win out'. The authors' previous report⁹ that females in this population favour paternal half-sisters over non-kin strengthens the interpretation that phenotypic cues are used to identify kin.

The likelihood of phenotype matching could be explored further by extending Buchan and colleagues' logistic regression approach. The authors partitioned juveniles into three discrete categories — offspring, unrelated but behaviourally predicted offspring, and non-offspring — but probability of paternity is actually a continuous variable. So, one could instead plot the regression between a male's probability of paternity (the proportion of a female's follicular phase that the male monopolized) and the relative frequency of male interventions. If the data points that fell above the regression line — the residuals — represented true (genetic) offspring significantly more often than non-kin, whereas points that fell below the line were the reverse, then this would indicate the occurrence of phenotype matching.

Regardless of the recognition mechanism, it is clear that male Amboseli baboons are fairly certain of their paternity. This casts doubt on the conventional wisdom that female primates copulate with several males to confuse paternity (to reduce infanticide and increase the pool of males that provide care for young). It also raises anew the question of why females typically mate so many times. Perhaps multiple mating enables females to compare males' abilities to stimulate them during sexual encounters¹⁰. Or it could enhance sperm competition, which

would benefit females if more competitive sperm contain 'better' genes, resulting in more vigorous offspring that are more likely to survive and reproduce^{11,12}.

Paul W. Sherman is in the Department of Neurobiology and Behavior, Mudd Hall, Cornell University, Ithaca, New York 14853, USA.
e-mail: pws6@cornell.edu

Bryan D. Neff is in the Department of Biology, University of Western Ontario, London, Ontario N6A 5B7, Canada.
e-mail: bneff@uwo.ca

1. Sherman, P. W., Reeve, H. K. & Pfennig, D. W. in *Behavioural Ecology: An Evolutionary Approach* 4th edn (eds Krebs, J. R. & Davies, N. B.) 69–96 (Oxford Univ. Press, 1997).
2. Buchan, J. C., Alberts, S. C., Silk, J. B. & Altmann, J.

Nature **425**, 179–181 (2003).

3. Hasselquist, D. & Sherman, P. W. *Behav. Ecol.* **12**, 457–466 (2001).
4. DeWoody, J. A. & Avise, J. C. *J. Hered.* **92**, 167–172 (2001).
5. Neff, B. D. & Sherman, P. W. *Proc. R. Soc. Lond. B* **269**, 1435–1441 (2002).
6. Neff, B. D. *Nature* **422**, 716–719 (2003).
7. Kempenaers, B. & Sheldon, B. C. *Anim. Behav.* **51**, 1165–1173 (1996).
8. Hauber, M. E. & Sherman, P. W. *Trends Neurosci.* **24**, 609–616 (2001).
9. Smith, K., Alberts, S. C. & Altmann, J. *Proc. R. Soc. Lond. B* **270**, 503–510 (2003).
10. Eberhard, W. G. *Female Control: Sexual Selection by Cryptic Female Choice* (Princeton Univ. Press, 1996).
11. Birkhead, T. W. & Möller, A. P. (eds) *Sperm Competition and Sexual Selection* (Academic, San Diego, 1998).
12. Johnsen, A., Andersen, V., Sunding, C. & Lifjeld, J. T. *Nature* **406**, 296–299 (2000).

Cosmochemistry

Inside the cosmic blender

Alex N. Halliday

The Solar System has a largely uniform isotopic composition but with tantalizing small variations. Geochemists are trying to ascertain the mechanisms and types of stars that produced this state of affairs.

Long before astronomers presented us with spectacular images of the material swirling around other stars, Pierre-Simon Laplace (1749–1827) proposed that the planets of our Solar System formed from a circumstellar disk. The scale and degree of isotopic heterogeneity of the matter in our Solar System provide unique insights into the dynamics associated with the development of these disks, and on page 152 of this issue Becker and Walker¹ highlight some of the problems associated with one class of the isotopes concerned — those of molybdenum.

Modern mass spectrometry techniques had revealed that the isotopic compositions of many of the more refractory elements in meteorites, including a primitive class of meteorite called chondrites, are, within error, identical to those found on Earth itself². It was unclear how the Solar System could have such a uniform composition without some kind of earlier mixing process, because the mechanism that produces these elements creates huge isotopic heterogeneity. The elements heavier than hydrogen and helium ('metals' to astronomers) are manufactured by fusion and irradiation in stars, in a process known as nucleosynthesis. Different kinds of mechanisms produce distinct isotopes, depending on the mass of the star, its stage of development and its metal composition. The most likely way to produce a uniform mix of atoms from the many stars that fed the molecular cloud that eventually formed our Solar System was for the building-blocks of the Earth and planets to have condensed from a hot, well-mixed gas. This idea led to a series of 'hot nebula' models, but various lines of evidence have since shown

that such models can only apply to localized portions of the Solar System.

The most important evidence has stemmed from the discovery of presolar grains³. A variety of grains condense in stellar envelopes and so record the extreme isotopic composition generated in the star itself. These have been discovered in chondrites — which are now viewed as having formed from relatively cold dust and debris in the circumstellar disk. The development of techniques for measuring the isotopic compositions of these submicrometre specks of stardust has produced data even for trace elements such as molybdenum and zirconium⁴. These remarkable samples of other stars have isotopic compositions that are completely unlike that of our Solar System. But they do match the compositions predicted from theory for various kinds of stars^{3,4}.

Some types of presolar grain could not have survived the high temperatures and chemistry advocated in the hot nebula model. Therefore, this stardust must have been introduced to a cold disk. There is no obvious mechanism for large-scale mixing in the molecular cloud that preceded disk formation, so such heterogeneity was probably eliminated by mixing in the disk itself, otherwise we could not explain the isotopic uniformity of the Solar System.

Astronomers have found evidence that disks act like swirling conveyor belts, accreting gas and dust onto the central star⁵. Most of the disk consists of hydrogen and helium, and its drag on the dust may result in lateral and radial mixing that eliminates variations⁶. Precise isotopic measurements could potentially allow small variations to be resolved and provide a map of mixing in the

disk at the start of our Solar System. Now, with improvements in mass spectrometric techniques, geochemists have indeed begun to report very small isotopic differences between bulk meteorites and planets in refractory elements that should have been predominantly inherited from the dust, such as chromium, zirconium and molybdenum. This is exciting, because unravelling such structure to the degree of mixing in the disk could provide unique insights into disk dynamics.

Some of the first evidence of radial variations in the dust came from chromium isotopes⁷. The interpretation of chromium-53 effects in terms of inherited or injected heterogeneity versus variations in radioactive manganese-53 caused by chemical factors has been a matter of debate⁷. Other chromium isotopic effects cannot be explained in this way, however. Early reports of large variations in zirconium isotopes have not been confirmed in more detailed studies, although there is a hint of extremely small degrees of heterogeneity⁸. Two groups^{9,10} have reported molybdenum isotopic variations, but their results were obtained using different methods, and are not the same. In their paper¹, Becker and Walker present the results of extensively replicated measurements on molybdenum, and disagree with the conclusions of the previous studies^{9,10}. They argue that no convincing isotopic differences can be resolved at the bulk rock scale, and that any residual effects left after correcting for instrumental factors probably represent artefacts that are poorly understood.

Becker and Walker express caution about their conclusions, which may well turn out to have been wise. Other research has left a wisp of smoke from a gun, and three reasons for further study.

First, the degree to which an isotopic anomaly can be resolved in some data sets seems to depend on the isotopes used to normalize the data for mass fractionation. This is true to varying degrees in the molybdenum results of all groups, although it is less apparent in the data of Becker and Walker. Such effects give the impression that subtle but real heterogeneities exist, even though they are hard to quantify. Unless there is some unknown complication, such as different mass discrimination laws for different samples and standards when measured in the laboratory, the most likely explanation is that there are small variations in atomic abundance between samples. Second, some of the patterns in the anomalies, when normalized with isotopes produced in the same nucleosynthetic pathways, are similar to those predicted from incomplete mixing of specific components from some stellar environments (such as certain types of red-giant stars or supernovae)^{9,10}. Again, this seems an unlikely coincidence. Third,

analyses of presolar grains⁴ and leached bulk chondrites⁸ reveal that huge anomalies consistent both with these stellar environments and with the claimed bulk rock anomalies are present at some scale in the samples.

Becker and Walker demonstrate that even with a number of very careful, well-replicated measurements it is premature to draw any firm conclusions about incomplete mixing from molybdenum isotopes at this stage. The bottom line is that we have a hint, but little more, of large-scale heterogeneity in the distribution of molybdenum isotopes in the disk. To resolve matters further we need more precise and reproducible methods, and must better characterize and correct the experimental artefacts that plague some attempts to make such measurements with confidence at the < 0.01% mark. It is to be expected that small isotopic anomalies exist at some level. In fact, they may well be easier to detect in other refractory elements that also have a range of isotopes generated by a variety of stellar processes.

It may not be long before we can relate specific anomalies to specific classes of meteorites, and hence their probable provenance in the Solar System. As astronomers provide ever-better definitions of stellar disks and their mixing environments, we will get a better idea of how to interpret meteorites and the source of their components. We may eventually end up with partially reconstructed maps showing the spatial distribution of the subtle variations in the relative propor-

tions of the nuclei mainly produced in supernovae versus those mainly produced in red giants, for example.

Ultimately, it would be desirable to map the Solar System's isotopic heterogeneity from the Sun to its outer reaches in the Oort cloud. Missions are planned that will provide a better idea of the composition of the solar wind, Mercury and comets. It would, of course, be nice to achieve a clear picture of how the 'cosmic blender' worked from data transmitted in such space missions. But the levels of precision needed to resolve the small effects in molybdenum and many other elements currently require mass spectrometers that are bulky and delicate, with considerable power requirements. Therefore, we are very unlikely to obtain information of the necessary precision without additional missions that return samples for analysis on Earth. ■

Alex N. Halliday is in the Department of Earth Sciences, ETH Zentrum, NO C61, Sonneggstrasse 5, Zurich CH-8092, Switzerland.

e-mail: halliday@erdw.ethz.ch

1. Becker, H. & Walker, R. J. *Nature* **425**, 152–155 (2003).
2. Suess, E. *Rev. Astron. Astrophys.* **3**, 217–234 (1965).
3. Nittler, L. R. *Earth Planet. Sci. Lett.* **209**, 259–273 (2003).
4. Nicolussi, G. K. et al. *Astrophys. J.* **504**, 492–499 (1998).
5. Hartmann, L. *Space Sci. Rev.* **92**, 55–68 (2000).
6. Weidenschilling, S. J. *Mon. Not. R. Astron. Soc.* **180**, 57–70 (1977).
7. Lugmair, G. W. & Shukolyukov, A. *Geochim. Cosmochim. Acta* **62**, 2863–2886 (1998).
8. Schönbachler, M. et al. *Earth Planet. Sci. Lett.* (in the press).
9. Yin, Q., Jacobsen, S. B. & Yamashita, K. *Nature* **415**, 881–883 (2002).
10. Dauphas, N., Marty, B. & Reisberg, L. *Astrophys. J.* **565**, 640–644 (2001).

Cardiovascular biology

Signalling silenced

Edward M. Conway and Peter Carmeliet

The mechanism by which the TIMP-2 protein inhibits blood-vessel formation has been uncovered — and it is not as expected. The finding has implications for treating a cancer by cutting off its blood supply.

Tumours and inflamed tissues cannot grow without blood vessels to supply them with oxygen. So interfering with the formation of new blood vessels has been proposed as an attractive means of combating both cancer and inflammation¹. One class of proteins that stimulates blood-vessel formation (angiogenesis) is the matrix metalloproteinases (MMPs), which are kept in check by molecules called tissue inhibitors of MMPs (TIMPs). TIMPs suppress both angiogenesis and tumour growth, suggesting that synthetic compounds mimicking the MMP-inhibiting effects of these proteins might prove useful in treating cancer. But such efforts have generally been disappointing^{1–4}. This could be explained by the fact that TIMPs also have other effects, and, writing in *Cell*, Seo et al.⁵ reveal a further, surprising function for one key TIMP. They

show that it prevents angiogenesis by blocking a cellular signalling pathway, and that this ability has nothing to do with its MMP-inhibiting capacity.

The idea that blocking angiogenesis might prove beneficial for treating cancer is supported by recent evidence that an inhibitor of a prime angiogenic molecule — vascular endothelial growth factor (VEGF) — prolongs survival of cancer patients (see ref. 6 for a comment on this finding). One mechanism by which VEGF promotes angiogenesis is by stimulating protein-degrading (proteolytic) enzymes, including MMPs^{1–4}. When first discovered, MMPs were thought to work mainly by breaking through the dense thicket of fibres that surrounds cells, thereby allowing cancer cells and blood vessels to invade surrounding tissues. The TIMPs, which inhibit the proteolytic activity

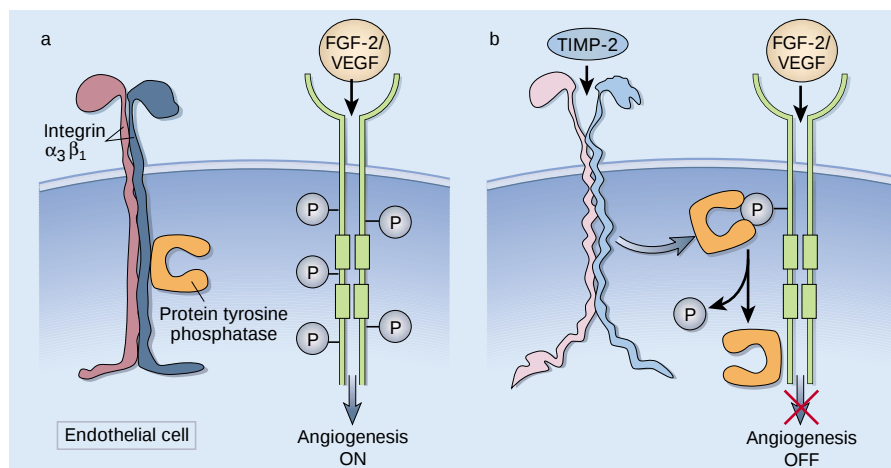


Figure 1 Blocking blood-vessel growth: how TIMP-2 does it. TIMP-2 inhibits matrix metalloproteinases (MMPs), and it has been assumed that it is this activity that predominantly underlies its ability to prevent blood-vessel growth (angiogenesis). But Seo *et al.*⁵ find otherwise. **a**, On endothelial cells (those that line blood vessels), when the receptors for vascular endothelial growth factor (VEGF) and fibroblast growth factor-2 (FGF-2) bind these molecules, they self-phosphorylate (P). This sets off a series of intracellular events that triggers angiogenesis. Here, $\alpha_3\beta_1$ integrin is associated with protein tyrosine phosphatases. **b**, When TIMP-2 binds to $\alpha_3\beta_1$ integrin, the phosphatases dissociate, and then reassociate with the VEGF and FGF-2 receptors. The phosphatases dephosphorylate the receptors, silencing them and thereby switching off angiogenesis.

of MMPs, suppress the migration of endothelial cells (which line blood vessels) and angiogenesis^{1–4}. Together, these findings provided the rationale for treating cancer with synthetic compounds that, like TIMPs, block MMP activity—a formidable effort in which the pharmaceutical industry has invested at least a billion dollars over the past 20 years.

But matters are rarely straightforward when it comes to living organisms. For starters, MMPs have a much broader spectrum of targets than originally believed, including other proteinases, proteinase inhibitors, latent growth factors, cell-surface receptors and clotting factors^{1–4}. Moreover, MMPs have diverse—and often opposing—effects on angiogenesis. For instance, they

can stimulate angiogenesis by cleaving matrix components and releasing growth factors that are sequestered in the matrix (such as VEGF), exposed on cell membranes or bound to precursors. But they suppress angiogenesis by processes that include activating cell-death pathways, generating angiogenesis inhibitors, and inactivating receptors for angiogenic molecules.

The dogma that TIMPs suppress angiogenesis only by blocking MMP activity has also been challenged, for example by the finding that a synthetic MMP inhibitor, batimastat, does not impair endothelial-cell growth⁷. And some unexpected roles for TIMPs have been discovered: for instance, TIMP-1 actually stimulates VEGF-driven angiogenesis, and TIMP-3 blocks angiogenesis by preventing VEGF from binding to its receptor⁸. All of this may help to explain why synthetic MMP inhibitors have not lived up to expectations, and have been plagued with undesirable side effects.

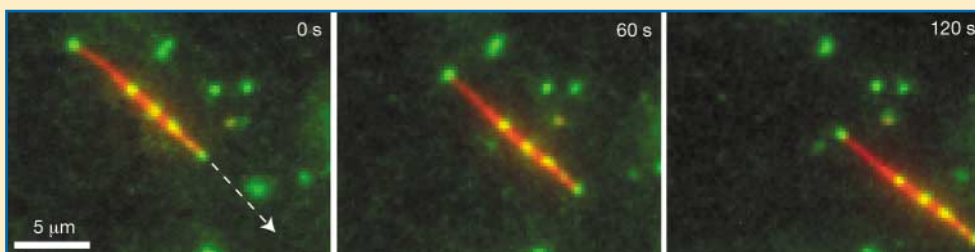
The TIMP studied by Seo *et al.*⁵ is TIMP-2, which is perhaps the most biologically relevant angiogenesis inhibitor of this family, as it most consistently blocks endothelial-cell migration and proliferation. How does it do this? More than a decade ago, Stetler-Stevenson *et al.*⁹ identified this protein as an inhibitor of MMP-mediated proteolysis; they later showed¹⁰ that transferring the TIMP-2 gene into tumours inhibits angiogenesis. At the time, it seemed likely that it does this through its effects on MMPs. But later studies showed that TIMP-2 can down-regulate VEGF expression¹¹. And recently,

Biophysics

Take the intracellular train

Proteins or neurotransmitters that are synthesized in one part of a cell yet needed in another usually catch the intracellular train: the molecules snuggle up inside vesicles that are then moved along microtubule tracks by motor proteins. Stefan Diez and colleagues have tapped into this transport system to mechanically manipulate DNA molecules and explore the potential of the cell's molecular transport machinery for engineering complex nanostructures and nanocircuits (*Nano Lett.* doi:10.1021/nl034504h; 2003).

In this work, biotin molecules at both ends of the DNA strands facilitate binding to streptavidin molecules, which are attached to the microtubules. When solutions of the microtubules are exposed to surfaces coated with kinesin motor proteins, the kinesin motors



attach to the DNA-carrying microtubules and slide them along the surface. This sequence of images shows five green-fluorescing DNA molecules, condensed into 1- μ m-wide coils, hitching a ride on a gliding red-fluorescing microtubule. When the pH of the solution is lowered, some DNA strands bind at one end to the surface; the other end can attach itself through the biotin–streptavidin linkage to a passing microtubule, and the DNA is stretched until it

detaches from the surface or breaks.

Using surfaces with high surface densities of kinesin, up to 100 motor molecules interact with a 5- μ m-long microtubule and power it along. But at much lower motor densities, the power output of individual kinesin motors barely suffices to stretch a coiled-up DNA strand: surface-anchored DNA can then act as a leash to stop microtubule movement or force it onto a circular path.

Diez *et al.* point out that

improved control might be achieved by using surfaces that have predefined gliding tracks to transport DNA between contact points in an array. By tailoring the sequence of the DNA molecules and employing restriction enzymes, unwanted connections could be cut out. Once such a system is in place, cargo other than DNA—such as functionalized carbon nanotubes—should be able to join the train. **Magdalena Helmer**

Fernandez *et al.*¹² found that TIMP-2 blocks the growth of endothelial cells in an MMP-independent way. This inhibition requires a peptide at the carboxy-terminal end of TIMP-2, which was superior to an amino-terminal TIMP-2 fragment (which blocks MMP activity) in inhibiting angiogenesis.

Now, Stetler-Stevenson's group (Seo *et al.*) brings further insight into the MMP-independent activity of TIMP-2. The authors show that TIMP-2 silences two growth-factor receptors — one that detects VEGF, and one for fibroblast growth factor-2 (FGF-2). They also offer clues to how this happens.

Many angiogenic factors, such as VEGF and FGF-2, bind to receptors on endothelial cells, causing the receptors to self-activate by adding phosphate groups to tyrosine amino acids in their intracellular portion¹³. This leads to a sequence of events that promotes endothelial-cell proliferation and migration, and hence angiogenesis. This process is partly regulated by enzymes that remove phosphate groups, such as SHP-1 and other 'protein tyrosine phosphatases', which prevent the receptors from transmitting further signals¹⁴.

Seo *et al.* show *in vitro* that a modified form of TIMP-2, which can no longer interact with MMPs, causes SHP-1 and other unidentified protein tyrosine phosphatases to associate with VEGF and FGF-2 receptors. This occurs in a remarkably indirect way: TIMP-2 binds to a neighbouring receptor on endothelial cells, called $\alpha_3\beta_1$ integrin, with which phosphatases usually associate. This binding causes the phosphatases to move from the integrin to the VEGF and FGF-2 receptors (Fig. 1). The cells are thereby rendered resistant to VEGF and FGF-2 — an effect that Seo *et al.* also observed in a mouse model of angiogenesis. The authors further show that inhibitors of protein tyrosine phosphatases largely block the anti-angiogenesis effects of TIMP-2. The discovery of such receptor crosstalk reveals a high degree of coordination between proteinase inhibitors, integrins and growth-factor receptors. It also provides insight into how signalling from VEGF receptors is terminated — a poorly understood process.

What are the medical implications of these exciting findings? As mentioned above, the development of synthetic MMP inhibitors for cancer treatment has been based, at least in part, on the observation that TIMP-2 and other TIMPs inhibit tumour growth in animal models¹⁻⁴. As a consequence, several MMP inhibitors have been extensively evaluated in clinical trials, but with disappointing results. But these inhibitors were selected for their TIMP-like ability to block the proteolytic activity of MMPs — and the findings of Seo *et al.*⁵ imply that the anti-angiogenic function of TIMP-2 stems primarily from its MMP-independent activity.

This conclusion needs to be confirmed. And it must be reconciled with gene-inactivation studies¹⁵ showing that TIMP-2's main function *in vivo* is to activate pro-angiogenic MMP-2, and with the finding that, for certain cancers, high levels of TIMP-2 predict poor (not, as one might expect for an angiogenesis inhibitor, good) prognosis. If Seo and colleagues' findings prove correct, however, it might be bad news for the strategy of using synthetic MMP inhibitors in the clinic. The good news, though, is that we now have other ideas for angiogenesis inhibitors: it might, for instance, be worth investigating analogues or fragments of TIMP-2 that induce phosphatases to silence the angiogenic growth-factor receptors. ■

Edward M. Conway and Peter Carmeliet are at the Centre for Transgene Technology and Gene Therapy, Flanders Interuniversity Institute for Biotechnology, University of Leuven, B-3000 Leuven, Belgium.

e-mail: peter.carmeliet@med.kuleuven.ac.be

1. Overall, C. M. & Lopez-Otin, C. *Nature Rev. Cancer* **2**, 657–672 (2002).
2. Brinckerhoff, C. E. & Matrisian, L. M. *Nature Rev. Mol. Cell Biol.* **3**, 207–214 (2002).
3. Egeblad, M. & Werb, Z. *Nature Rev. Cancer* **2**, 161–174 (2002).
4. Coussens, L. M., Fingleton, B. & Matrisian, L. M. *Science* **295**, 2387–2392 (2002).
5. Seo, D. W. *et al. Cell* **114**, 171–180 (2003).
6. McCarthy, M. *Lancet* **361**, 1959 (2003).
7. Murphy, A. N., Unsworth, E. J. & Stetler-Stevenson, W. G. *J. Cell Physiol.* **157**, 351–358 (1993).
8. Qi, J. H. *et al. Nature Med.* **9**, 407–415 (2003).
9. Stetler-Stevenson, W. G., Krutzsch, H. C. & Liotta, L. A. *Matrix* **1** (Suppl.), 299–306 (1992).
10. Valente, P. *et al. Int. J. Cancer* **75**, 246–253 (1998).
11. Hajitou, A. *et al. Cancer Res.* **61**, 3450–3457 (2001).
12. Fernandez, C. A., Butterfield, C., Jackson, G. & Moses, M. A. *J. Biol. Chem.* published online 4 August 2003 (doi:10.1074/jbc.M306176200).
13. Schlessinger, J. *Cell* **103**, 211–225 (2000).
14. Ostman, A. & Bohmer, F. D. *Trends Cell Biol.* **11**, 258–266 (2001).
15. Wang, Z., Juttermann, R. & Soloway, P. D. *J. Biol. Chem.* **275**, 26411–26415 (2000).

Clarification

Readers may have been misled by the phrasing of the penultimate paragraph in David Gems and Joshua J. McElwee's News and Views article "Ageing: Microarraying mortality" (*Nature* **424**, 259–261; 2003). The article dealt with a paper by C. T. Murphy *et al.* in the same issue (**424**, 277–284; 2003), which described the use of microarrays and RNA interference to identify *Caenorhabditis elegans* longevity genes that are regulated by insulin/IGF-1 signalling. The phrase concerned was "Seeking insight into the biochemistry of ageing, Murphy *et al.* panned their mine of microarray data for genes that fit existing expectations — an approach sometimes referred to as 'fishing'". This did not refer to the choice of genes for functional investigation with RNAi, which, with the exception of *ins-7*, involved testing clones with the greatest expression change and/or the highest overall statistical significance. Rather, the comment was directed to the discussion of the biochemical significance of the overall data.



100 YEARS AGO

The Scottish Sanitary Congress was opened at Stranraer on Thursday last... Prof. Glaister, in the course of his remarks, urged that men of science and local authorities should realise the detrimental effect of atmospheric pollution, and together grapple with the subject. The prejudicial effects of town living could not be better demonstrated than in the depreciated physique of the third and fourth generations of many of those who had proceeded from the country to the towns. One of the significant features of present-day statistics, and one calling for the serious consideration of sanitarians, was the high prevailing rate of infantile mortality in populous centres. If the state of the principal English towns for 1901 be considered, it will be found that the infantile death rate varied from 126 per thousand up to 226 per thousand. These figures exhibited a great wastage of infantile life. He affirmed that it was a preventable wastage, and, therefore, worthy the reflections of sanitarians. Such high rates of infantile mortality were bound in the future to become a serious national concern in view of the diminution of the birth rate which had been progressively taking place for the last few decades. From *Nature* 10 September 1903.

50 YEARS AGO

In an imaginative but well-authenticated article, Nigel Balchin, the well-known psychologist and author, examines the way in which the growth of industry has affected the outlook and philosophy of those who work in large or small industrial undertakings... A comparison is made between the life of the agricultural worker and of the individual whose working life is confined to the factory, and Balchin concludes that the impact of industry upon the worker has been to destroy the simple and direct significance of work done and hence to sever the mental link between 'work' and life. In these circumstances 'work' is liable to become something which has no logic or point in itself. The impact of the huge and complicated industry structure on the worker has been to confuse the mind as to his real desires and requirements and to leave him with a vague impression that man exists for industry rather than *vice versa*.

From *Nature* 12 September 1953.

Obituary

Thomas Risley Odhiambo (1931–2003)

In 1965, a short communication, 'Metabolic effects of corpus allatum hormone, in the desert locust, *Schistocerca gregaria*', appeared in *Nature*. Thomas R. Odhiambo was the sole author, and this paper marked his arrival in the field of insect physiology. Odhiambo died of liver cancer in Nairobi, Kenya, on 26 May this year. His legacy lies not only in his work as an entomologist, but in the enduring achievements that stemmed from his promotion of scientific development across Africa.

Odhiambo completed his studies in botany and zoology at Makakere University College, Kampala, Uganda, in 1953. After graduating, he spent three years at the Tea Research Institute in Kericho, Kenya, before pursuing his fascination with African insects at the Severe and Kawanda research stations in Uganda. There, he was responsible for the discovery of several new genera and dozens of new species of the Miridae bugs (Hemiptera) of East Africa.

In 1959, he moved to study at the University of Cambridge. He spent the next six years completing, first, an MA in natural sciences, and then a PhD under the supervision of the guru of insect physiology, Vincent Wigglesworth. It is said that Wigglesworth claimed that Odhiambo was the best student he had ever had. Certainly, Odhiambo's thesis was phenomenally productive — a series of 14 papers on the reproductive physiology of the desert locust stemmed from it.

There was expectation about the fruits of his next labours, and we did not have long to wait for them. In 1966 he published his research on the feeding behaviour and reproductive physiology of the tsetse fly, which he had begun in his new position as lecturer at the University of East Africa in Nairobi. His selection of insects was not arbitrary; the desert locust and the tsetse fly are major African pests, and Odhiambo continued to study them throughout his career.

But his research into these insects did not consume all of his considerable energies. Since his student days in Makekere, Odhiambo had continually agonized over the status of science in Africa. He knew that, historically, Africans had been world leaders in mathematics, astronomy and engineering in ancient Egypt, and also in Cartaga, Timbuktu and Zimbabwe. Yet, after the fifteenth century, African scientific enterprise seemed to have evaporated. Odhiambo was convinced that if the right environmental conditions



Champion of scientific development in Africa — and author of children's books

were fostered, Africa's native potential would re-emerge. He lamented the increasing brain drain from Africa, because talented students, receiving fellowships to study abroad, were reluctant to return home where they could not apply their newly acquired skills.

Odhiambo believed that scientists such as himself could be instrumental in changing the scientific conditions across the entire African continent. In 1967 he was approached by *Science* to write a review on the status of science in Africa. In this article Odhiambo made a plea for the establishment of several large centres of excellence in Africa. He proposed that these centres should be staffed by world authorities in specific scientific disciplines, who should mentor bright young postgraduate and postdoctoral fellows from Africa and other developing countries, to build the first generation of African scientists.

Strong support for Odhiambo's ideas came from, among others, Carl Djerassi in the United States. Together they set the wheels in motion for launching the International Center of Insect Physiology and Endocrinology (later changed to Ecology), ICIPE, in Nairobi. They approached several foreign academies and research institutes for help, and in 1969 ICIPE was established. One of the objectives of the new centre was to create motivated and highly talented 'human capital' in insect research and related areas of science, so as, to quote Odhiambo, "to enable Africa to sustain herself and to lead

the entire pan-tropical world in this area of endeavour".

Eventually, 21 national academies became sponsors of ICIPE, and they provided the external research directors. Odhiambo was the principal director, and he successfully headed the centre for 25 years. Running ICIPE was quite an art. The elite participating institutions expected the centre to devote itself largely to basic science, whereas the donors pushed for applications of the research. Odhiambo's charisma, intelligence, diplomacy and patience enabled him to navigate his ship successfully through these tricky waters. In the early years, the external directors supervised the research. But both Odhiambo and the governing board understood that the ultimate proof of ICIPE's success would be its Africanization. Today, the ICIPE remains a centre of scientific excellence and training. And as Odhiambo had hoped, it is indeed staffed mainly by African scientists.

Odhiambo's aspirations for scientific development in Africa were larger still — he worked tirelessly to increase the priority of science in the policies of African nations. To this end, he helped to establish the Third World Academy in 1985, and he founded the African Academy of Sciences in 1987, which he chaired until 1999. He also recognized that scientific publication is an integral part of scientific research and that African scientists need to be heard. As a result, he was instrumental in creating the Academy Science Publishers and ICIPE Science Press, and launching the two journals *Insect Science and its Application* and *Discovery and Innovation*. Odhiambo also understood the compelling reasons for introducing children to science at an early age. He established ChiSci Scientific Publications for this purpose, and he himself wrote six science books for children.

In March this year, ICIPE honoured Thomas Odhiambo for his vision and his dedication to creating a science-based infrastructure in Africa that could provide the solutions to a multitude of problems in the tropics. He will be sorely missed by his friends and colleagues the world over. Rachel Galun and Onesmo K. ole-MoiYoi
Rachel Galun is in the Department of Parasitology, Hebrew University of Jerusalem, Jerusalem 91905, Israel.
e-mail: zoher@netvision.net.il
Onesmo K. ole-MoiYoi is at ICIPE, Duguville, PO Box 30772-00100, Nairobi, Kenya.
e-mail: oole-moiyoi@icipe.org

Ecology

Human past and forest present

J. Biogeogr. **30**, 1381–1390 (2003)

Human impacts on tropical rainforests are not necessarily confined to the past few decades, say Barend S. van Gernerden and colleagues. They conclude that the composition of tree species in a present-day African forest is partly a legacy of agriculture several centuries ago.

Different tree species prefer different conditions in their youth. Some are good at establishing themselves in small gaps, such as those created when a single tree falls; others prefer the gaps created by large disturbances — storms or fire, for instance, or human activities. The researchers surveyed 16 one-hectare forest plots in Cameroon. In nine of them, the oldest, biggest trees were known to prefer large-scale disturbance.

The size of these trees suggested that they were 300–400 years old, and the high level of charcoal found in the soil implied that the land was once used by humans — fires are rare in these forests. Van Gernerden and colleagues argue that this view of ecological history should be incorporated into thinking about what constitutes 'pristine forest' and so into conservation planning. **John Whitfield**

Oceanography

No go with the floe

Geophys. Res. Lett. doi:10.1029/2003GL017721 (2003)

C-19 was born in May 2002. This was not an unromantically named baby but a huge iceberg, 32 km wide and 200 km long, that calved from the western Ross Ice Shelf in Antarctica. Kevin R. Arrigo and Gert L. van Dijken used satellite data to track its stately progress through the Ross Sea and, following its grounding near the passage leading to the open ocean, to assess the consequences.

In 'normal ice years', much of the southwestern Ross Sea is clear of sea ice during the austral spring and summer, and the conditions are ideal for the proliferation of phytoplankton. In the 2002–03 season, however, ice piled up behind C-19, reducing the open-water area to about a quarter of the norm. Phytoplankton growth, as measured by chlorophyll concentrations, was ten times less than usual.

An elder sibling of C-19, iceberg B-15, calved and broke up in the Ross Sea in 2000. Although it was even bigger than C-19, its impact on phytoplankton was much less. It is not just size that matters — Arrigo and van Dijken point out that the

time and place of iceberg entry into the Ross Sea are crucial determinants of the ecological outcome.

Tim Lincoln

Neurobiology

The eyes have it

Cell **114**, 545–557 (2003)¹

Neuron **39**, 919–935 (2003)²

What do people, mice and frogs have but adult birds, fish and tadpoles lack? The answer is binocular vision: the ability to produce just one mental representation of images seen, despite receiving input from two eyes. According to Carol A. Mason and colleagues¹, this useful trick relies on a protein called Zic2.

In animals that use binocular vision, each side of the brain must receive information from both eyes. So some neurons (retinal ganglion cells) from the right eye must send projections that cross over the brain's midline into the left hemisphere. Others must project to the same side. Such 'uncrossed' projections are absent in animals that lack binocular vision.

Mason and colleagues now find that Zic2 is expressed in these uncrossed neurons during mouse development — and that when Zic2 is inactivated, there are few or no uncrossed projections. Moreover, animals such as ferrets, which have a higher degree of binocularity and more uncrossed projections, show greater Zic2 expression in the relevant region of the embryonic retina. So this protein seems to determine whether a ganglion cell will cross the midline. As Zic2 is a transcription factor, the next step is to see what genes it might be regulating, and the same group suggests² that the EphB1 gene is a candidate. **Amanda Tromans**

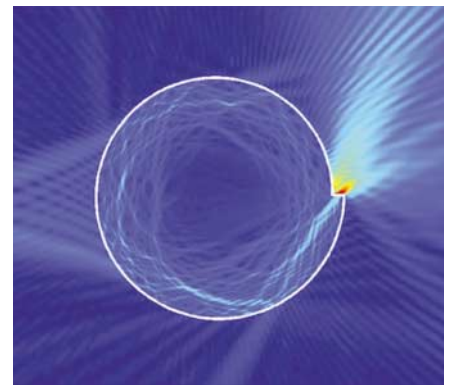
Microphotonics

Tiny laser notches up success

Appl. Phys. Lett. **83**, 1710–1712 (2003)

Microscopic lasers made from semiconductors can be used for photonic and optoelectronic information processing on a chip. Pillar-shaped microlasers can be created from multiple layers of different materials, like a stack of coins. The laser light is stimulated within 'quantum wells', which are formed by thin films of active material sandwiched between layers of another material. It all works nicely, except for the fact that most microlasers with this design produce at least two output beams, whereas a single, unidirectional laser beam is needed to arrange easy optical communication between microlasers.

G. D. Chern *et al.* have now found a way to achieve this in pillar-shaped, quantum-well microlasers. The trick is to carve each layer into the shape of a spiral, rather



The pillar-shaped microlaser. The colours indicate the varying strength of electric field inside the microlaser cavity. The field is strongest around the notch in the outside surface, from which a single laser beam can emerge.

like the silhouette of a snail shell. The laser light is excited in quantum wells of indium gallium nitride — but in a perfect disk, the light would merely rotate around the well without escaping. The 'notch' in the perimeter of the spiral-shaped layers provides a point of escape. When stimulated with ultraviolet light, the microlaser produces a single violet beam that emerges at an angle of 40° from the normal to the notch.

Philip Ball

Genomics

DNA double agent

Curr. Biol. **13**, 1518–1523 (2003)

Selfish DNA is genetic information that has inserted itself into our genome to serve its own purposes. In some cases this happened many millions of years ago. Back then, selfish DNA was active and could infect new cells. Nowadays, it is largely inactive and has decayed into a fossil-like state.

Clare Lynch and Michael Tristem have now found one of the most ancient selfish genetic elements described to date. Known as a gypsy LTR-retrotransposon, it has been hiding in the genomes of mice, rats, sheep and humans for more than 70 million years. The element appears to have lost sequences that would allow it to function as a retrotransposon. Yet, contrary to expectations, it is still active in other respects: it contains long regions of functional, coding DNA.

Lynch and Tristem propose that, over time, the retrotransposon has become domesticated by its hosts, and now serves a useful purpose. This may be the key to its long life. Although its exact function is unknown, the authors speculate that this genetic element may be a molecular double agent, guarding cells against infection by other members of its own family. **Helen R. Pilcher**

Directed bending of a polymer film by light

Miniaturizing a simple photomechanical system could expand its range of applications.

Polymer solutions and solids that contain light-sensitive molecules can undergo photo-contraction, whereby light energy is converted into mechanical energy^{1–8}. Here we show that a single film of a liquid-crystal network containing an azobenzene chromophore can be repeatedly and precisely bent along any chosen direction by using linearly polarized light. This striking photomechanical effect results from a photoselective volume contraction and may be useful in the development of high-speed actuators for microscale or nanoscale applications, for example in micro-robots in medicine or optical microtweezers.

We prepared the films by thermal polymerization of a liquid-crystal monomer (molecule **1** in Fig. 1a) and a diacrylate crosslinker (molecule **2**) (9:1, mol/mol), both of which possess azobenzene moieties. Figure 1b shows a sequence of frames indicating how such a film can be bent along a precisely controlled direction by irradiation with linearly polarized light (for movies, see supplementary information).

The first frame in Fig. 1b shows the film before light irradiation; counting clockwise, the second shows how the film curls up after exposure to 366-nm light that has a polarization direction at zero degrees. The film is bent towards the direction of irradiation of the light, with the bending occurring parallel to the direction of light polarization (Fig. 1b, white arrows). When the bent film was exposed to visible light with a wavelength longer than 540 nm, it completely reverted to its initial flat state (third frame in Fig. 1b).

The effects on the film of altering the polarization direction of light at 366 nm to -45° , -90° and -135° are shown in the fourth, sixth and eighth frames, respectively, of Fig. 1b. It can be seen that the bending direction of the film moves anticlockwise by 45° , 90° and 135° , respectively, keeping parallel to the direction of light polarization. The film could be restored from each bent state to its initial flat form by irradiation with visible light at wavelengths longer than 540 nm (Fig. 1b). Moreover, the bending–unbending cycle of these four modes could be repeated without apparent fatigue. These results show that the bending direction of a single film can be precisely controlled by altering the direction of polarization of the irradiating light, and that such a film can be bent repeatedly.

In the film used here, the incident light is mostly absorbed by the film surface because of the strong absorption by the azobenzene moieties of light at about 360 nm. The film is a polydomain liquid-crystal film that consists of many micro-sized domains of

azobenzene liquid-crystal moieties aligned in one direction in each domain, although macroscopically the direction of alignment is random. On irradiation of the film with linearly polarized light, the selective absorption of light of a specific direction leads to a *trans*–*cis* isomerization of the azobenzene moieties in specific domains where the azobenzene moieties are aligned along the direction of light polarization.

As a result, the subtle reduction in microscopic size and ordering of the liquid-crystal components^{1–4,9,10} gives rise to a substantial macroscopic volume contraction^{5–8} at the film surface and a bending of the whole film as a result of the cooperative movement of the liquid-crystal moieties and polymer segments. Light energy is therefore efficiently converted into mechanical energy. It should be simple to miniaturize our system for potential application in driving micro-machines and nanomachines without the

aid of batteries, motors and gears, using remote irradiation with laser beams.

Yanlei Yu, Makoto Nakano, Tomiki Ikeda

Chemical Resources Laboratory, Tokyo Institute of Technology, 4259 Nagatsuta, Midori-ku, Yokohama 226-8503, Japan

e-mail: tiked@res.titech.ac.jp

1. Van der Veen, G. & Prins, W. *Nature Phys. Sci.* **230**, 70–72 (1971).
2. Smets, G. & De Blauwe, F. *Pure Appl. Chem.* **39**, 225–238 (1974).
3. Eisenbach, C. D. *Polymer* **21**, 1175–1179 (1980).
4. Matejka, L., Ilavsky, M., Dusek, K. & Wichterle, O. *Polymer* **22**, 1511–1515 (1981).
5. Finkelmann, H., Nishikawa, E., Pereira, G. G. & Warner, M. *Phys. Rev. Lett.* **87**, 015501–1–4 (2001).
6. Hogan, P. M., Tajbakhsh, A. R. & Terentjev, E. M. *Phys. Rev. E* **65**, 041720–1–10 (2002).
7. Li, M.-H. et al. *Adv. Mater.* **15**, 569–572 (2003).
8. Ikeda, T., Nakano, M., Yu, Y., Tsutsumi, O. & Kanazawa, A. *Adv. Mater.* **15**, 201–205 (2003).
9. Ikeda, T. J. *Mater. Chem.* **13**, 2037–2057 (2000).
10. Ikeda, T., & Tsutsumi, O. *Science* **268**, 1873–1875 (1995).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

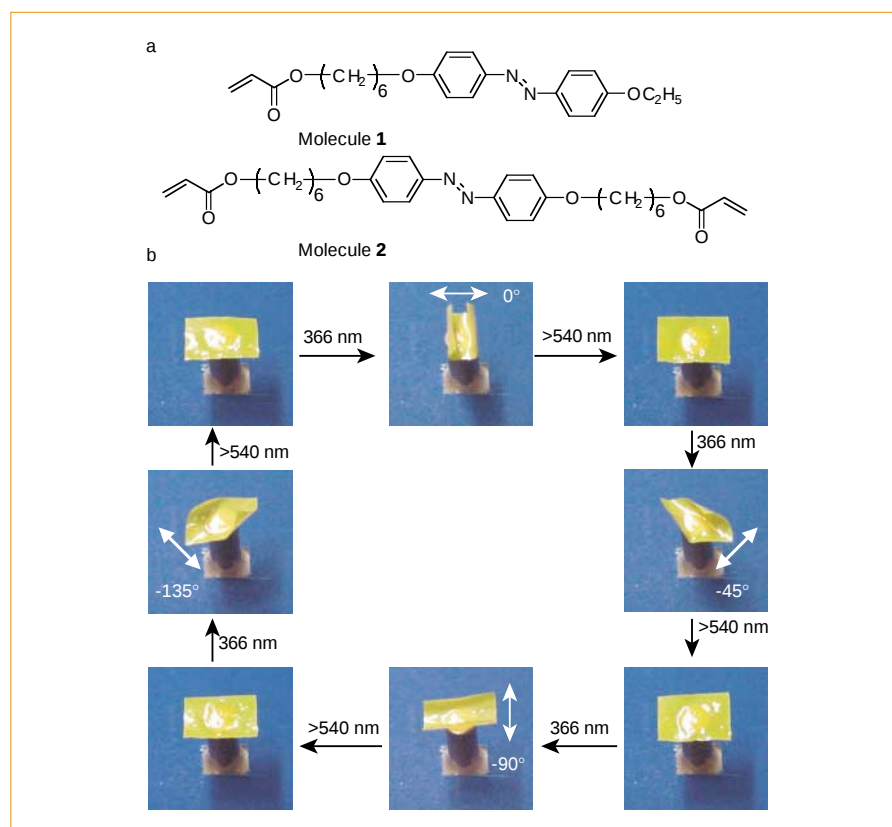


Figure 1 Precise control of the bending direction of a film by linearly polarized light. **a**, Chemical structures of the liquid-crystal monomer (molecule **1**) and crosslinker (molecule **2**) used for preparation of the film. On cooling from 95 °C, **1** changes from an isotropic to a nematic phase, and at 91 °C it becomes crystalline; molecule **2** undergoes a phase transition from crystalline to isotropic at 98 °C. **b**, Photographic frames of the film bending in different directions in response to irradiation by linearly polarized light of different angles of polarization (white arrows) at 366 nm, and being flattened again by visible light longer than 540 nm. The flat film (4.5 mm × 3 mm × 7 μm) lay on a copper stick fixed to a copper plate; a stage under the plate was set at 85 °C to control the temperature of the film, which was covered by a piece of blue paper. The bending time for the four different bending directions was within 10 s, when the light intensity of 366-nm linearly polarized light was 3.5 mW cm⁻² after exposure to visible light longer than 540 nm (547 nm, 24.2 mW cm⁻²; 577 nm, 26.8 mW cm⁻²), the bent film reverted to the flat state in about 10 s.

Reproduction

Widespread cloning in echinoderm larvae

Asexual reproduction by free-living invertebrate larvae is a rare and enigmatic phenomenon and, although it is known to occur in sea stars^{1–4} and brittle stars^{5,6}, it has not been detected in other echinoderms despite more than a century of intensive study^{6,7}. Here we describe spontaneous larval cloning in three species from two more echinoderm classes: a sea cucumber (Holothuroidea), a sand dollar and a sea urchin (Echinoidea). Larval cloning may therefore be an ancient ability of echinoderms and possibly of deuterostomes — the group that includes echinoderms, acorn worms, sea squirts and vertebrates.

To confirm that genuine cloning was occurring, we reared cloning larvae and separated clones individually (Fig. 1, legend). Although most sea cucumber larvae metamorphosed normally after a month, 5 of 41 (12.2%) free-swimming doliolaria larvae were half the length of their siblings and had a constriction around the penultimate ciliary band. These buds retained a ciliary band and were still attached by a thin tether into the benthic pentacula stage 12 h later (Fig. 1a). Eventually, after separation, buds developed into normal auricularia larvae with a juvenile rudiment (Fig. 1b).

Most sand dollar larvae metamorphosed within 7 weeks, but 6 of 170 (3.5%) formed a hollow bud near the future juvenile mouth (results not shown). Once separated, the evenly ciliated buds developed into an almost solid gastrula with bilateral spicules. Within two days, a tripartite gut and further skeletal spicules formed, and the clones began to feed.

After 4 weeks, nearly 5% of sea urchin larvae ($n > 500$) showed constrictions at their posterior end (Fig. 1c). These eventually yielded uniformly ciliated buds that began feeding within four days, developed paired skeletal rods within a week, and acquired a normal larval form within two weeks. Juvenile rudiments began to form within three weeks (Fig. 1d), accompanied by further larval arms.

Larval cloning is therefore known to occur in all classes except crinoids (feather stars and sea lilies), supporting an earlier conjecture that it might be an ancestral ability of echinoderms⁸. The mechanisms by which this cloning occurs, however, are unexpectedly diverse. First, clones may arise from various larval body regions, including arms (sea stars^{1–4}, brittle stars⁵), the oral hood (sea stars^{2–4}), the posterior end (sea stars³, sea cucumbers (Fig. 1a), sea urchins (Fig. 1c)) and the lateral body wall (sea stars², sand dollars (our results, not shown)). Second, the

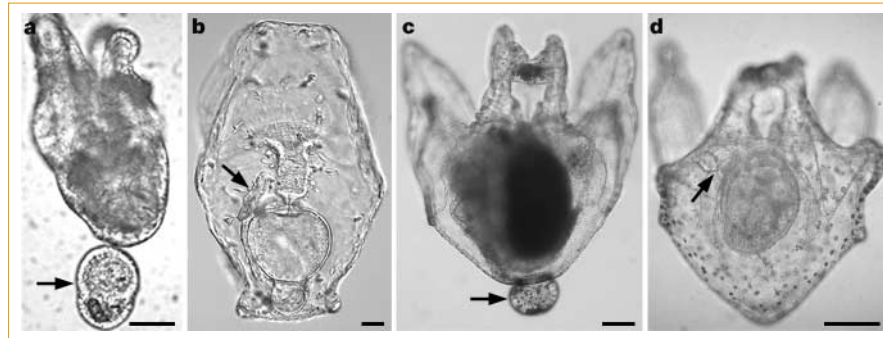


Figure 1 Asexual budding by echinoderm larvae. **a**, Sea cucumber (*Parastichopus californicus*; Holothuroidea; Aspidochirotrida) early pentacula with clone attached at posterior end (arrowhead). **b**, Sea cucumber 18 days after separation as a complete auricularia larva, with juvenile rudiment (arrow). **c**, Sea urchin (*Strongylocentrotus purpuratus*; Echinoidea, Echinacea) larva with a clone (arrow) constricting from the posterior end around the robust ciliary band. **d**, Sea urchin clone three weeks after separation, with early juvenile rudiment (arrow). Scale bars, 50 μ m. Adult *P. californicus* were collected sub-tidally from San Juan Island, Washington, and gametes were collected by dissection. Larvae were maintained at 11 °C under natural lighting in filtered sea water and were fed a red alga *Rhodomonas salina*, ad libitum. Adult *S. purpuratus* from Prasiola Point (Vancouver Island, Canada) were spawned by intracoelomic injection of 0.55 M potassium chloride. Larvae were maintained in a 19-h light cycle at 14 °C in pasteurized sea water and were fed $> 10^6$ cells per ml of a golden brown alga *Isochrysis galbana* and *R. salina*; food and water were replaced every 72 h. To monitor their development, cloning larvae and separated clones were isolated in 7-ml glass vials and cultured; most developed to metamorphosis. Cultures were deemed to be healthy as no larval tissue necrosis was seen. Further details are available from the authors.

developmental stage of clones at separation ranges from blastulae² to fully formed larvae³. Third, some clones may not separate until after the primary larva has begun to metamorphose⁵ (Fig. 1a). This indicates either that larval cloning evolved independently on several occasions or that its mechanisms have diverged widely from an ancestral mode.

Cloning can be surprisingly frequent: up to 12% in laboratory-reared sea cucumber larvae and 10–90% in samples of field-collected sea star larvae^{1,4}. The fact that such a common phenomenon should have been overlooked in heavily studied organisms seems remarkable. Were preconceptions about ‘normal’ development so strong that cloning was simply dismissed as aberrant? From a theoretical perspective, larval cloning — particularly in sea urchins — challenges a central tenet of the ‘set-aside’ cell theory⁹ because, contrary to prediction, larval body cells are not “essentially eutelic”⁹, but can differentiate into juvenile structures (Fig. 1b, d). Nevertheless, such cloning offers an opportunity to study the well-characterized developmental regulatory networks of sea urchins¹⁰ in a new ontogenetic context.

Larval cloning represents an intriguing new dimension to invertebrate life histories. The process confers three potential ecological advantages: increased fecundity under optimal growth conditions^{2,3}, increased chances of settlement after a protracted larval life^{1,2,5}, and recycling of otherwise discarded or reabsorbed larval tissue⁵. Larval cloning may also be evolutionarily significant, for two reasons. First, clones may subsequently clone⁵, potentially leading to a new, entirely pelagic bauplan. Second, although it may be merely another

idiosyncrasy of echinoderm development, larval cloning might also be more taxonomically widespread. As nearest relatives to the echinoderms¹¹, acorn worms offer a critical test. If their tornaria larvae clone, then ancient deuterostomes may have had this ability.

Alexandra A. Eaves*, A. Richard Palmer†

*Physiology and Cell Biology Group, and †Systematics and Evolution Group, Department of Biological Sciences, University of Alberta, Edmonton, Alberta T6G 2E9, Canada, and Bamfield Marine Sciences Centre, Bamfield, British Columbia V0R 1B0, Canada

e-mail: aeaves@ualberta.ca

1. Bosh, I., Rivkin, R. B. & Alexander, S. P. *Nature* **337**, 169–170 (1989).
2. Jaekle, W. B. *Biol. Bull.* **186**, 62–71 (1994).
3. Vickery, M. S. & McClintock, J. B. *Biol. Bull.* **199**, 298–304 (2000).
4. Knott, K. E., Balser, E. J., Jaekle, W. B. & Wray, G. A. *Biol. Bull.* **204**, 246–255 (2003).
5. Balser, E. J. *Biol. Bull.* **194**, 187–193 (1998).
6. Mortensen, T. *Studies of the Development and Larval Forms of Echinoderms* (Gad, Copenhagen, 1921).
7. MacBride, E. W. *Nature* **108**, 529–530 (1921).
8. Lacalli, T. C. *Invertebr. Biol.* **119**, 234–241 (2000).
9. Davidson, E. H., Cameron, R. A. & Ransick, A. *Development* **125**, 3269–3290 (1998).
10. Davidson, E. H. et al. *Science* **295**, 1669–1678 (2002).
11. Cameron, C. B., Garey, J. R. & Swalla, B. J. *Proc. Natl Acad. Sci. USA* **97**, 4469–4474 (2000).

Competing financial interests: declared none.

brief communications is intended to provide a forum for brief, topical reports of general scientific interest and for technical discussion of recently published material of particular interest to non-specialist readers (**communications arising**). Priority will be given to contributions that have fewer than 500 words, 10 references and only one figure. Detailed guidelines are available on Nature’s website (www.nature.com/nature).

The vector alignments of asteroid spins by thermal torques

David Vokrouhlický¹, David Nesvorný² & William F. Bottke²

¹Institute of Astronomy, Charles University, V Holešovičkách 2, 180 00 Prague 8, Czech Republic

²Southwest Research Institute, 1050 Walnut St, Suite 400, Boulder, Colorado 80302, USA

Collisions have been thought to be the dominant process altering asteroid rotations, but recent observations of the Koronis family of asteroids suggest that this may be incorrect. This group of asteroids was formed in a catastrophic collision several billion years ago; in the intervening period their rotational axes should have become nearly random because of subsequent collisions, with spin rates that follow a maxwellian distribution. What is seen, however, is that the observed family members with prograde spins have nearly identical periods (7.5–9.5 h) and obliquities between 42 and 50 degrees, while those with retrograde spins have obliquities between 154 and 169 degrees with periods either <5 h or >13 h. Here we show that these non-random orientations and spin rates can be explained by ‘thermal torques’ (arising from differential solar heating), which modify the spin states over time. In some cases, the asteroids become trapped in spin-orbit resonances. Our results suggest that thermal torques may be more important than collisions in changing the spin states (and possibly shapes) of asteroids with diameters <40 km.

Asteroid families are by-products of catastrophic disruption events (for example, ref. 1). They are identified by the clustered proper semimajor axis a , eccentricity e , and inclination I values of their members^{2,3} as well as their common spectral properties⁴. Because these traits also allow us to estimate the ages of different families, family members can be used to investigate, in a constrained manner, collisional and dynamical evolution in the main belt over the last few Gyr. In contrast, non-family asteroids, which have a poorly constrained origin, are more difficult to use for this purpose.

We focus on the Koronis asteroid family, whose members are located at proper $a = 2.83\text{--}2.95$ AU, $e = 0.04\text{--}0.09$, and $\sin I = 0.032\text{--}0.042$ (ref. 5). Numerical simulations suggest that the Koronis family members were produced when an asteroid of diameter $D \approx 60$ km struck the $D \approx 120$ km Koronis parent body at 3 km s^{-1} (ref. 6). The approximate age of this family, based on studies of craters on asteroid (243) Ida, a member of the Koronis family observed by the Galileo spacecraft, and by collisional/dynamical evolution simulations, is 2–3 Gyr (refs 5, 7–9). Our calculations suggest that most $20 < D < 40$ km asteroids in this family have not suffered a disruption event over this time¹⁰. If the heavily cratered surface of (243) Ida is any guide¹¹, however, these asteroids have experienced numerous non-disruptive impact events since their formation.

Insights gleaned from laboratory and numerical experiments of asteroid collisions suggest that Koronis family members should have spin rates that approximately follow a maxwellian distribution and nearly random spin-axis orientations^{6,12,13}. The observed rotation states of 20–40 km diameter asteroids within this family, however, are surprisingly different from these results. Led by clues that Koronis family members may have peculiar rotation rates and preferentially aligned spin vectors (which was thought to indicate that the Koronis family was young)^{14,15}, Slivan repeatedly observed ten Koronis family objects, including (243) Ida, for nearly a decade^{16,17}. Information on these objects can be found in Table 1. Slivan found that the prograde rotators had tightly clustered values in spin period ($7.5 < P < 9.5$ h), obliquity ($42^\circ < \epsilon < 50^\circ$), and possibly ecliptic longitude, with the latter two implying that the spin axes of this group are truly parallel in space. We will refer to prograde objects with spin vectors near these clustered values as being in ‘Slivan states’. Retrograde rotators, on the other hand, had $P < 5$ h or $P > 13$ h, $\epsilon \geq 154^\circ$, and ecliptic longitudes that appear to span a large range of values.

Slivan’s solutions have been verified in two different ways. First of all, his analysis yielded a spin vector for (243) Ida that agrees with spacecraft observations^{17,18}. Second, Slivan’s results have been reproduced via an independent analysis performed¹⁷ using the methods described in refs 19 and 20. Hence, the solutions described in Table 1 appear to be robust.

Given the singular nature of the spin vectors in both the prograde and retrograde groups, we consider it extremely unlikely that collisions alone could have produced this distribution; some other explanation is needed. Here we show that the observed spin vectors are a by-product of dynamical processes that act over long time intervals. If true, these same processes should also play an important role in modifying the spin states of nearly all main belt asteroids with $D < 40$ km. Note that other than the results of refs 16 and 17, the spin states of main belt asteroids in this size range have been little explored.

Spin vector evolution model

We constructed a numerical model that tracks the evolution of an asteroid’s rotation state over 4 Gyr via two processes: (1) thermal torques produced by the reflection and re-emission of sunlight from an asteroid’s surface (that is, the Yarkovsky–O’Keefe–Radzievskii–Paddack effect or YORP effect)^{21–23}, and (2) the effect of solar

Table 1 Spin vector information for Koronis family asteroids

Asteroid	D (km)	P (h)	ϵ (°)	λ_1 (°)	λ_2 (°)
(311) Claudia	24	7.53	50	24	209
(534) Nassovia	34	9.47	42	55	241
(720) Bohlinia	34	8.92	50	48	236
(1223) Neckar	22	7.82	47	73	259
(158) Koronis	36	14.21	159	27	211
(167) Urda	40	13.06	163	39	225
(208) Lacrimosa	42	14.08	156	162	346
(243) Ida	28	4.63	156	–	263
(277) Elvira	28	29.69	169	51	242
(321) Florentina	28	2.87	154	94	265

Columns included mean asteroid diameter D , sidereal rotation period P , adopted mean obliquity ϵ , and two solutions for the ecliptic longitude λ , with $0^\circ < \lambda_1 \leq 180^\circ$ and $180^\circ < \lambda_2 \leq 360^\circ$ (refs 16–18). The diameter is computed from the absolute magnitude and geometric albedo estimated for each object. The uncertainty in ϵ is $\pm 5^\circ$ and the uncertainty in λ is $\pm (5^\circ\text{--}20^\circ)$. The first four asteroids rotate in a prograde sense ($\epsilon < 90^\circ$), while the last six asteroids rotate in a retrograde sense ($\epsilon > 90^\circ$). We also point out that an additional Koronis family asteroid, (462) Eriphyla, has recently been observed, and its size ($D = 38$ km), spin period ($P = 8.66$ h), spin-axis obliquity ($\epsilon = 57^\circ$), and longitude ($\lambda_1 = 103^\circ$) are all consistent with the clustered values in the prograde group (S. Slivan & M. Kaasalainen, personal communication).

gravitational torques on an asteroid whose orbit changes over time due to planetary gravitational perturbations²⁴. The YORP effect is related to the so-called Yarkovsky effect, a thermal radiation force that causes objects to undergo semimajor axis drift as a function of their spin, orbit, size and material properties²³. YORP torques, although they are also dependent on these factors, are additionally affected by an object's precise shape; energy re-radiated from an irregularly shaped body (such as a propeller or windmill) allows the YORP effect to change its spin rate and obliquity over time, while energy re-radiated from a symmetrical body (such as a sphere or ellipsoid) produces no net YORP torque^{21–23}.

To determine how the spin-vector evolution of Koronis family members is affected by shape, we input real and artificial asteroid configurations into our model (such as the spacecraft derived shape of (243) Ida and numerous random asteroid shapes derived using numerical techniques^{22,25,26}). Our results agree with previous work: the spin states of irregular shapes evolve more quickly via the YORP effect than rounded shapes and the evolutionary timescale is proportional to D^2 (refs 21–23). An example of this would be the spin-vector evolution of (243) Ida. If we input into our simulations an Ida shape model derived from Galileo spacecraft images²⁷, we find it evolves twice as fast as a more-rounded shape derived from lightcurve-inversion techniques^{17,19,20}. Nevertheless, in each case, the spin-vector endstate reached by the test asteroid was the same;

only the timescale was different. Hence, in the runs described below, we used asteroid shapes that were realistic and which conformed to the performance expected from highly accurate shape models.

Prograde rotators captured by spin-orbit resonances

Using our model, we explored the long-term evolution of Koronis family asteroids having prograde and retrograde spins. For prograde rotators, the results relevant to the observed Koronis family members in Table 1 are those test asteroids that predominately spin down. Figure 1 shows the representative evolution of (311) Claudia, which was started with $P = 5$ h and a range of ϵ values. We find that the YORP torque affecting obliquity, defined as τ_{obl} , slowly drives ϵ toward the 0° asymptotic value and, in the process, increases $\dot{\psi}$, defined as the asteroid's precession rate with respect to the ecliptic. This occurs because $\dot{\psi} \propto \cos \epsilon / \omega$, where ω is the asteroid's rotation rate. When the initial ϵ is small, or if it becomes so in the course of the evolution, the YORP torque affecting P , defined as τ_{spin} , decelerates the rotation rate and therefore also increases the precession rate of the spin axis until the motion becomes captured in a spin-orbit resonance. In the case of the Koronis family asteroids, the first important resonance encountered is with Saturn's longitude of node (defined for low-enough inclination by $\dot{\psi} \approx -s_6$, where s_6 is the mean frequency of Saturn's longitude of node, or $-26.34'' \text{ y}^{-1}$; refs 28, 29). We find the capture probability is unity for ϵ smaller than $\approx 30^\circ$ (see also ref. 30).

Once in a spin-orbit resonance, the body's spin and resonance frame precess at the same rate, which keeps the spin-axis longitude fixed with respect to a specific value determined by the resonance while YORP torques modify the body's P and ϵ values. This explains why the prograde asteroids in Table 1 have clustered longitude values (see also Fig. 2). Note that there is a 180° ambiguity in the ecliptic longitude solutions presented in Table 1 (that is, λ_1 and λ_2).

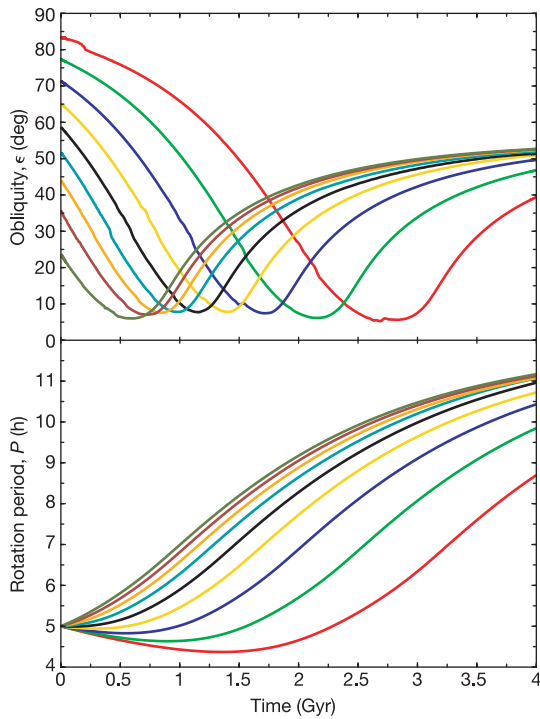


Figure 1 Several possible evolutionary paths for prograde rotator (311) Claudia, many of which evolve into Silvan states. We set Claudia's initial spin period to $P = 5$ h and its initial obliquity ϵ to uniform values in $\cos \epsilon$ between 0° – 90° . We find that YORP torques modify P while driving ϵ to small values. Eventually, our test asteroids are captured by the s_6 spin-orbit secular resonance between the precession rate of the asteroid's spin axis and Saturn's longitude of node. This occurs near the minimum ϵ values for each curve shown above. Once trapped in the resonance, P steadily increases, enough to cause migration of its equilibrium (Cassini) state^{30,41} and force ϵ to move toward the asymptotic value of $\approx 55^\circ$. We find that varying the shapes of our asteroids mainly affects the rate at which P and ϵ evolve rather than their endstate, with elongated, asymmetrical bodies evolving to the equilibrium point faster than rounded bodies. To make the figure easier to read, short-period variations in obliquity have been filtered out. Their amplitude is typically 5° – 10° (especially in the resonant phase), so that all solutions, except the one starting with $\epsilon \approx 84.3^\circ$, are consistent with observations.

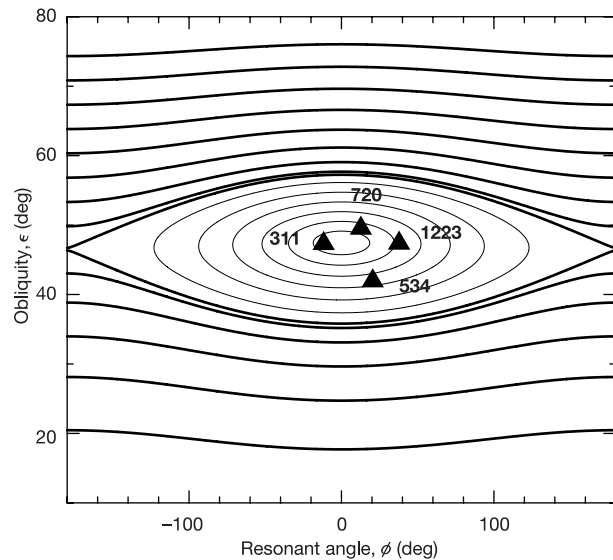


Figure 2 Prograde Koronis family members trapped in the s_6 spin-orbit resonance. This resonance is defined by $\dot{\psi} \approx -s_6$, with $\dot{\psi}$ the precession rate and $s_6 = -26.34'' \text{ y}^{-1}$ (the mean frequency of Saturn's longitude of node). Solid lines show the libration and circulation paths. Here we chose $\alpha/s_6 = -1.47$ and $\tau_{\text{obl}} = 0$ (see the Methods section). The abscissa is the resonant angle $\phi = -(\dot{\psi} + s_6 t + \Phi_6)$, with $\dot{\psi}$ the longitude of the spin vector, t being time, and Φ_6 being the phase of the frequency term s_6 . We find $\Phi_6 \approx 307.3^\circ$. The ordinate is the obliquity ϵ . Stable and unstable points are the Cassini states^{30,41}; in this case the stable point has obliquity $\epsilon = 47^\circ$, consistent with bodies that have evolved for 2–3 Gyr but have not yet reached their 'endstate' ($\epsilon \approx 55^\circ$). Symbols show projections of the prograde rotators from Table 1. Our solutions for these values use the ecliptic longitude λ_1 . The prograde asteroids are located well inside the libration zone of the resonance, implying they are currently trapped in the resonance.

As we show below, our model results predict that λ_1 is the true solution. Unfortunately, ground-based observations may be unable to resolve this issue in the near future (S. Slivan, personal communication).

As τ_{spin} continues to decelerate ω , ϵ increases to maintain the resonant condition $\dot{\psi} \approx -s_6$. Eventually, this process forces ϵ to converge to $\approx 55^\circ$, a value where the YORP torque τ_{spin} becomes zero for nearly all asteroid shapes²². At present, this surprising but important result has only been numerically determined. The $\approx 55^\circ$ value can probably be traced to the fact that a sphere with the same obliquity has a yearly-averaged insolation that is independent of latitude²⁸. This property allows the YORP effect to drive prograde asteroids towards Slivan states (see Methods).

Depending on their shape and the particular spin-orbit resonance that produces the trapping event, asteroids in Slivan states may have a variety of P values. For example, in some of our runs, test asteroids trapped in spin-orbit resonances other than s_6 have $P > 15$ h. The prograde asteroids in Table 1, however, apparently had similar enough sizes, shapes and initial P values that all reached a Slivan state in the s_6 spin-orbit resonance after 2–3 Gyr of evolution.

We find that test asteroids with initial $4 < P < 7$ h yield results similar to the Fig. 1 case. Test asteroids with $P > 7$ h, however, have initial ϵ (and $\dot{\psi}$) values that either cause them to jump or place them beyond the s_6 spin-orbit resonance. Only slow rotators started with high ϵ values have an opportunity to reach spin states like those in Table 1. The rest may eventually reach Slivan states by becoming trapped in other spin-orbit resonances, but their P values would be higher than those observed. For $P < 4$ h, we find YORP torques

cannot drain rotational angular momentum out of the test asteroids fast enough for them to reach Slivan states within 4 Gyr. Thus, it is plausible that the Koronis family-forming event produced numerous $20 < D < 40$ fragments with P values originally between 4–7 h.

To check our results, we tested whether the prograde rotators were in a resonant state by projecting their observed spin vectors onto the plane of the appropriate resonant variables (Fig. 2). We find that their spin vectors cluster near the centre of the libration zone of the s_6 resonance, exactly where they should be if they are trapped in that resonance. This resonance-locking mechanism is what produces the clustered λ_1 longitude values in Table 1.

Our results indicate that (311) Claudia and (720) Bohlinia may be more elongated or asymmetrical than is suggested by their lightcurve-derived shapes¹⁷. We do not believe this is a problem because the methods used¹⁷ to derive these shapes may round asteroid ‘hull’ dimensions by as much as 30%. Moreover, (311) Claudia exhibits lightcurve behaviour that is very similar to Ida’s and cannot be fully explained by ellipsoidal models (S. Slivan, personal communication). An intriguing alternative explanation is that their spin-axis precession rates were modified over time by the presence of a kilometer-sized satellite. Note that we already know that at least one Koronis family member, (243) Ida, has a small satellite³¹.

Note that all spin-orbit equilibrium points are weakly unstable and may cause asteroid spin vectors to leave the resonance after some time period (see ref. 22 for examples). This effect may have already happened to small Koronis family members (for example, $D < 10$ km), whose ϵ and P values evolve quickly via the YORP effect^{21,22}. To regain a Slivan state, an asteroid needs to become trapped in another spin-orbit resonance.

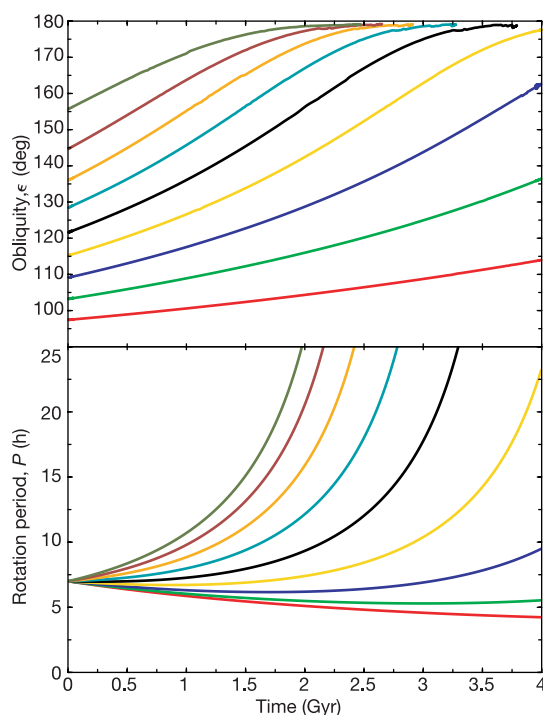


Figure 3 Several possible evolutionary paths for retrograde rotator (167) Urda, many of which evolve to slower rotation rates over 4 Gyr. We set its initial spin period to $P = 7$ h and its initial obliquity ϵ to uniform values in $\cos \epsilon$ between 90° – 180° (the short-period oscillations were eliminated as in Fig. 1). In most cases, YORP torques drive the obliquity ϵ towards the asymptotic value of 180° while increasing the object’s rotation period P . In the remaining cases, ϵ slowly increases while P decreases. Note that there are no meaningful spin-orbit resonances for retrograde objects, so these evolutionary paths are very different from those shown in Fig. 1. The black curve provides a good match to the present state of Urda after 2–3 Gyr of evolution. The other curves provide insights into the probable evolutionary paths followed by the remaining objects in the retrograde group (Table 1).

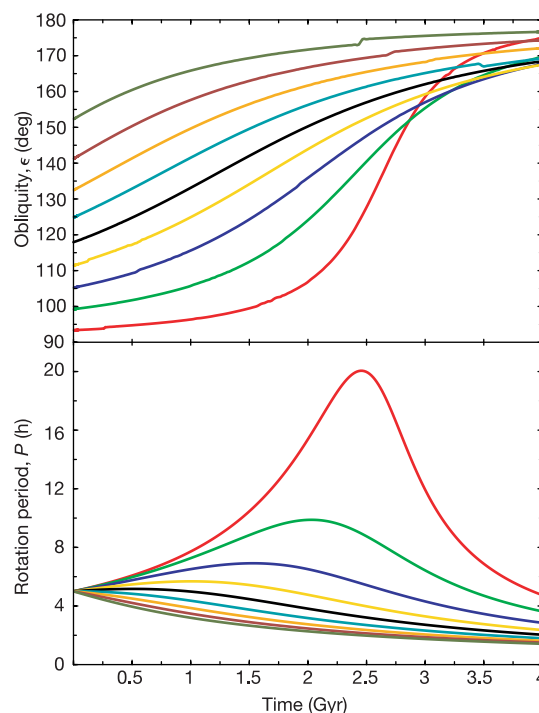


Figure 4 Several possible evolutionary paths for retrograde rotator (321) Florentina, many of which evolve to faster rotation rates. Florentina’s initial spin period was set to $P = 5$ h and its initial obliquity ϵ was set uniformly in $\cos \epsilon$ between 90° – 180° . Like the results in Fig. 3, YORP torques drive the obliquity ϵ towards the asymptotic value of 180° . In this case, however, they decrease rather than increase the object’s rotation period P . The black curve provides a good match to the present state of Florentina after 2–3 Gyr of evolution.

We also investigated prograde rotators that predominately spin up rather than down. These asteroids are unlikely to be captured by the s_6 spin-orbit resonance and therefore are inconsistent with Table 1 data. For this reason, we do not discuss them further here.

Spin-vector endstates for retrograde rotators

Our model shows that retrograde rotators have their obliquity values steadily driven by the YORP effect toward the asymptotic value of $\epsilon = 180^\circ$. By varying the initial P , ϵ values of our test asteroids, and by giving them different shapes, we found a variety of possible evolutionary paths. Some test asteroids always spin down, while others always spin up. There were even cases where objects start by spinning down but then later spin up or vice versa (Figs 3 and 4). Nevertheless, all these evolutionary paths inevitably push retrograde rotators toward the extreme P values found in Table 1.

Figure 3 shows the representative evolution of (167) Urda that was given an initial $P = 7$ h and a variety of obliquity values. We chose parameters that caused our test asteroid to predominately spin down. Because there are no important spin-orbit resonances for retrograde rotators, except for those with $P \geq 60$ h, the spin states in Fig. 3 change slowly as ϵ is driven towards 180° . Gravitational planetary perturbations play a minimal role in the evolution of these objects, such that clustered longitude values are not expected. Interestingly, Table 1 shows five out of six asteroids with λ solutions within $\sim 50^\circ$ of one another. Unlike the prograde case, however, we have no reason to think λ_1 or λ_2 is preferred.

Our results show numerous combinations of initial P , ϵ values that reproduce the spin state of (167) Urda after 2–3 Gyr of evolution. For example, depending on the ϵ value selected, initial $6 < P < 15$ h values yield reasonable matches. The only cases that do not appear to work for Urda are test asteroids with $P < 6$ h. YORP torques generally cannot eliminate enough rotational angular momentum from these bodies over 4 Gyr to match Urda's constraints.

Fast retrograde rotators like (243) Ida and (321) Florentina probably had initial shapes that forced them to predominately spin up. Figure 4 shows the representative evolution of (321) Florentina, which had observed $P = 2.87$ h. As in the Urda case, our results show many positive matches for a wide range of initial P , ϵ values. If Florentina continues along its current evolutionary path, and if it is a gravitational aggregate, it will probably shed mass in order to remove excess rotational angular momentum. Assuming that rotational disruption is analogous to tidal disruption of an asteroid/comet near a planet^{32,33}, this event may create a small asteroid satellite. We speculate that Dactyl, the satellite of (243) Ida, may have been produced by just such an event in the past.

Implications and predictions

We conclude that the YORP effect, working in concert with planetary perturbations, provides the most plausible means of explaining the spin rate and obliquity values of the Koronis family asteroids described in Table 1. Moreover, if the observed prograde rotators are actually trapped in the s_6 spin-orbit resonance, as indicated by our results, it would imply that collisions have played a minimal role in the evolution of ϵ and P over the last 2–3 Gyr; otherwise, we would expect that s_6 locking would have been interrupted by non-disruptive collisions. A possible explanation is that relatively little rotational angular momentum is transferred from the projectile to the target asteroid during a cratering event (for example, ref. 34).

More generally, these results suggest that, over the last several Gyr, the YORP effect may have been more efficient than collisions at changing the spin rates and obliquities of main-belt asteroids with $D < 40$ km. If true, YORP can provide a natural explanation for the plethora of $D < 40$ km asteroids with extremely fast or slow rotation rates³⁵. We caution, however, that many small asteroids do not have extreme P values, implying that collisions do play some role in the evolution of their spin states. Understanding the complex

interplay between collisional effects, YORP, and semimajor axis drift via the Yarkovsky effect may be critical to determining how asteroids (and meteoroids) evolve throughout the inner Solar System^{36,37}.

The YORP effect may also have far-reaching implications for the physical history of many small asteroids. If asteroids are predominantly gravitational aggregates (or 'rubble-pile' asteroids³⁸), YORP may spin up some of them so fast that they change shape, shed mass, or even undergo fission. It may thus provide an important means of producing asteroid satellites among smaller main-belt and near-Earth asteroids (for example, timescales for spinning up kilometer-sized asteroids beyond their rotational break-up limit are a few Myr; refs 21–23).

To investigate whether asteroids in other main-belt regions may also be trapped in Slivan states, we tracked the spin-vector evolution of test prograde rotators having the same orbital parameters as (8) Flora, (15) Eunomia, (20) Massalia, (24) Themis, (37) Fides, and (221) Eos. These asteroids were chosen because they span the main belt in both semimajor axis and inclination. Our preliminary results indicate that low inclination asteroids in the outer main belt (such as (24) Themis) follow evolutionary paths similar to those described in Fig. 1; we predict that many more Slivan-state asteroids may be found there. For the rest of our test asteroids, we found more complex spin-vector evolutionary paths brought on, in part, by the presence of overlapping spin-orbit resonances. Resonance trapping events in these regions frequently force ϵ to undergo huge oscillations rather than migrating to $\approx 55^\circ$ (refs 22, 24, 39). Hence, to gain new insights into the history of inner main-belt asteroids or those with high inclinations, it is important first to make a careful comparison between numerical results and observations. $\square$

Methods

Because we are interested in long-term asteroid evolution via the YORP effect, short timescales, such as the asteroid's proper rotation or its revolution around the Sun, are eliminated by averaging. The evolving parameters include the spin rate ω (rotation about the shortest axis of the inertia tensor is assumed), obliquity ϵ and precession in longitude ψ (see ref. 29 for their definition). We use the complex variable $\xi = \sin \epsilon \exp(i\psi)$ to eliminate the apparent singularity at $\epsilon = 0^\circ$ and $\epsilon = 180^\circ$. The problem is then described by²²:

$$\frac{d\omega}{dt} = \tau_{\text{spin}} \quad (1)$$

$$\frac{d\xi}{dt} = i\xi(\alpha \sqrt{1 - \xi\bar{\xi}} - 2C) \pm (\mathcal{A} + i\mathcal{B})\sqrt{1 - \xi\bar{\xi}} \mp \frac{\tau_{\text{obl}}\xi}{\omega} \sqrt{\frac{1}{\xi\bar{\xi}} - 1} \quad (2)$$

with the precession constant $\alpha = 1.5 n^2 \Delta / \omega$, n is the asteroid mean motion around the Sun, $\bar{\xi}$ is the complex conjugate quantity, and $\Delta = 1 - (A + B)/(2C)$ with A , B and C the principal moments of the inertia tensor. Upper signs are for $0^\circ \leq \epsilon < 90^\circ$ and lower signs are for $90^\circ < \epsilon \leq 180^\circ$. We define τ_{spin} as the component of the YORP torque affecting ω and τ_{obl} the component affecting ϵ . Note that the solar torque is much more effective at determining the asteroid's precession rate than YORP. The effect of the orbital plane variations is included through the functions²⁹:

$$\mathcal{A} + i\mathcal{B} = \frac{2}{\sqrt{1 - \xi\bar{\xi}}} \left(\frac{d\xi}{dt} - i\xi C \right) \quad (3)$$

$$C = \frac{1}{2i} \left(\bar{\xi} \frac{d\xi}{dt} - \xi \frac{d\bar{\xi}}{dt} \right) \quad (4)$$

where $\xi = \sin I/2 \exp(i\Omega)$ with the orbital inclination I and longitude of node Ω . The $\mathcal{A}$, $\mathcal{B}$, and C functions for all asteroids in our sample (Table 1) were obtained by direct numerical integration and an accurate representation with Fourier series⁴⁰. The most important terms in this series for the Koronis family have proper and forced frequencies of $s \approx -67'' \text{ y}^{-1}$ and $s_6 \approx -26.34'' \text{ y}^{-1}$, respectively; the latter has a phase $\Phi_6 \approx 307.3^\circ$ with time origin at J2000.0.

When limited to a single periodic term, the conservative problem without YORP torques is integrable and often referred to as Colombo's top^{40,41}; its equilibrium solutions are called Cassini states. In the generalized case described by equations (1) and (2), which contains the YORP effect, the equilibrium solution still exists but it now constrains obliquity to $\approx 55^\circ$ where $\tau_{\text{spin}} = 0$ (see text). In this case, equation (2) sets the equilibrium value of P . Performing a linear-perturbation analysis, we find that, unlike in the solely gravitational case, the equilibrium solution with YORP is unstable.

The YORP effect in equations (1) and (2) is represented by projections:

$$\tau_{\text{spin}} = \frac{\mathbf{T} \cdot \mathbf{e}}{C}, \quad \tau_{\text{obl}} = \frac{\mathbf{T} \cdot \mathbf{e}_\perp}{C} \quad (5)$$

of the thermal torque $\mathbf{T}$ onto the spin unit vector $\mathbf{e}$ and the perpendicular vector $\mathbf{e}_\perp = [\mathbf{n} - (\mathbf{n} \cdot \mathbf{e})\mathbf{e}] / \sin \epsilon$ with $\mathbf{n}$ the unit normal to the orbital plane. Averaging over an asteroid's rotation and revolution around the Sun is assumed in equation (5). To account for shape

effects, we also average equation (5) over a large sample of gaussian random spheres^{25,26} with results generalized to include the effect of finite surface conductivity (values between 0.001 to 0.01 W m⁻¹ K⁻¹ are used). We find that, unlike in the zero-conductivity case²², τ_{obl} drives the ϵ value of most asteroids to the asymptotic values of 0° or 180° (see also ref. 42). When ϵ is near these values, τ_{spin} may be either positive or negative, corresponding to asymptotically decelerating or accelerating the rotation rate, respectively. For low-conductivity surfaces, the former case appears to be slightly more likely for the asteroid configurations tested.

Received 27 May; accepted 28 July 2003; doi:10.1038/nature01948.

- Zappalà, V., Cellino, A., Dell'Oro, A. & Paolicchi, P. in *Asteroids III* (eds Bottke, W. F. *et al.*) 619–631 (Univ. Arizona Press, Tucson, 2003).
- Knežević, Z., Lemaître, A. & Milani, A. in *Asteroids III* (eds Bottke, W. F. *et al.*) 603–612 (Univ. Arizona Press, Tucson, 2003).
- Bendjoya, P. & Zappalà, V. in *Asteroids III* (ed. Bottke, W. F. *et al.*) 613–618 (Univ. Arizona Press, Tucson, 2003).
- Cellino, A., Bus, S. J., Doressoundiram, A. & Lazzaro, D. in *Asteroids III* (eds Bottke, W. F. *et al.*) 632–643 (Univ. Arizona Press, Tucson, 2003).
- Bottke, W. F., Vokrouhlický, D., Brož, M., Nesvorný, D. & Morbidelli, A. Dynamical spreading of asteroid families by the Yarkovsky effect. *Science* **294**, 1693–1696 (2001).
- Michel, P., Benz, W., Tanga, P. & Richardson, D. C. Collisions and gravitational reaccumulation: Forming asteroid families and satellites. *Science* **294**, 1696–1700 (2001).
- Chapman, C. R. in *Asteroids III* (eds Bottke, W. F. *et al.*) 315–329 (Univ. Arizona Press, Tucson, 2003).
- Marzari, F., Davis, D. & Vanzani, V. Collisional evolution of asteroid families. *Icarus* **113**, 168–187 (1995).
- Greenberg, R. *et al.* Collisional and dynamical history of Ida. *Icarus* **120**, 106–118 (1996).
- Bottke, W. F. & Greenberg, R. in *Collisional Processes in the Solar System* (eds Marov, M. Y. & Rickman, H.) 51–71 (Kluwer Academic, Dordrecht, 2001).
- Belton, M. J. S. *et al.* First images of asteroid 243 Ida. *Science* **265**, 1543–1547 (1994).
- Asphaug, E. & Scheeres, D. J. Deconstructing Castalia: Evaluating a postimpact state. *Icarus* **139**, 383–386 (1999).
- Holsapple, K., Giblin, I., Housen, K., Nakamura, A. & Ryan, E. in *Asteroids III* (eds Bottke, W. F. *et al.*) 443–462 (Univ. Arizona Press, Tucson, 2003).
- Binzel, R. P. Collisional evolution in the Eos and Koronis asteroid families—Observational and numerical results. *Icarus* **73**, 303–313 (1988).
- Binzel, R. P. in *Asteroids, Comets, Meteors III*, (eds Lagerkvist, C. I., Richman, H. & Lindblad, B. A.) 15–18, (Uppsala, 1990).
- Slivan, S. M. Spin vector alignment of Koronis family asteroids. *Nature* **419**, 49–51 (2002).
- Slivan, S. M. *et al.* Spin vectors in the Koronis family: Comprehensive results from two independent analysis of 213 rotation lightcurves. *Icarus* **162**, 285–307 (2003).
- Davies, M. E., Colvin, T. R., Belton, M. J. S., Veverka, J. & Thomas, P. C. The direction of the north pole and the control network of asteroid 243 Ida. *Icarus* **120**, 33–37 (1996).
- Kaasalainen, M. & Torppa, J. Optimization methods for asteroid lightcurve inversion. I. Shape determination. *Icarus* **153**, 24–36 (2001).
- Kaasalainen, M., Torppa, J. & Muinonen, K. Optimization methods for asteroid lightcurve inversion. II. The complete inversion problem. *Icarus* **153**, 37–51 (2001).
- Rubincam, D. P. Radiative spin-up and spin-down of small asteroids. *Icarus* **148**, 2–11 (2000).
- Vokrouhlický, D. & Čapek, D. YORP-induced long-term evolution of the spin state of small asteroids and meteoroids. Rubincam's approximation. *Icarus* **159**, 449–467 (2002).

- Bottke, W. F., Vokrouhlický, D., Rubincam, D. P. & Brož, M. in *Asteroids III* (eds Bottke, W. F. *et al.*) 395–408 (Univ. Arizona Press, Tucson, 2003).
- Sköglöv, E., Magnusson, P. & Dahlgren, M. Evolution of the obliquities for ten asteroids. *Planet. Space Sci.* **44**, 1177–1183 (1996).
- Muinonen, K. Introducing the Gaussian shape hypothesis for asteroids and comets. *Astron. Astrophys.* **332**, 1087–1098 (1998).
- Muinonen, K. & Lagerros, J. S. V. Inversion of shape statistics for small solar system bodies. *Astron. Astrophys.* **333**, 753–761 (1998).
- Thomas, P. C. *et al.* The shape of Ida. *Icarus* **120**, 20–32 (1996).
- Ward, W. R. Climatic variations on Mars. Astronomical theory of insolation. *J. Geophys. Res.* **79**, 3375–3386 (1974).
- Laskar, J. & Robutel, P. The chaotic obliquity of the planets. *Nature* **361**, 608–612 (1993).
- Henrard, J. & Murigande, C. Colombo's top. *Celest. Mech.* **40**, 345–366 (1987).
- Belton, M. J. S. *et al.* The discovery and orbit of 1993 (243)1 Dactyl. *Icarus* **120**, 185–199 (1996).
- Richardson, D. C., Bottke, W. F. & Love, S. G. Tidal distortion and disruption of Earth-crossing asteroids. *Icarus* **134**, 47–76 (1998).
- Asphaug, E. & Benz, W. Size, density, and structure of comet Shoemaker-Levy 9 inferred from the physics of tidal breakup. *Icarus* **121**, 225–248 (1996).
- Love, S. G. & Ahrens, T. J. Origin of asteroid rotation rates in catastrophic impacts. *Nature* **386**, 154–156 (1997).
- Pravec, P., Harris, A. W. & Michałowski, T. in *Asteroids III* (eds Bottke, W. F. *et al.*) 113–122 (Univ. Arizona Press, Tucson, 2003).
- Vokrouhlický, D. & Farinella, P. Efficient delivery of meteorites to the Earth from a wide range of asteroid parent bodies. *Nature* **407**, 606–608 (2000).
- Morbidelli, A. & Vokrouhlický, D. The Yarkovsky-driven origin of near Earth asteroids. *Icarus* **163**, 120–134 (2003).
- Richardson, D. C., Leinhardt, Z. M., Melosh, H. J., Bottke, W. F. & Asphaug, E. in *Asteroids III* (eds Bottke, W. F. *et al.*) 501–515 (Univ. Arizona Press, Tucson, 2003).
- Rubincam, D. P., Rowlands, D. D. & Ray, R. D. Is asteroid 951 Gaspra in a resonant obliquity state with its spin increasing due to YORP? *J. Geophys. Res.* **107**(E9), 5065 (2002).
- Šidlichovský, M. & Nesvorný, D. Frequency modified Fourier transform and its applications to asteroids. *Celest. Mech. Dyn. Astron.* **65**, 137–148 (1997).
- Colombo, G. Cassini's second and third laws. *Astron. J.* **71**, 891–896 (1966).
- Čapek, D. & Vokrouhlický, D. in 'Asteroids, Comets and Meteors' Conf. (Berlin, 2002) abstract 12-07, 117–118; at (<http://berlinadmin.dlr.de/SGF/acm2002/>).

Acknowledgements We thank C. Agnor, D. Čapek, C. Chapman, D. Durda, M. Kaasalainen, A. Morbidelli, S. Slivan and W. Ward for discussions and comments. We also thank R. Binzel and D. Rubincam for reviews of our manuscript. The project was supported by NASA's Planetary Geology and Geophysics programme, NRC's COBASE programme, and the Grant Agency of the Czech Republic.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.V. (vokrouhl@mbox.cesnet.cz).

Efficient mixing of the solar nebula from uniform Mo isotopic composition of meteorites

Harry Becker & Richard J. Walker

Department of Geology, University of Maryland, College Park, Maryland 20742, USA

The abundances of elements and their isotopes in our Galaxy show wide variations, reflecting different nucleosynthetic processes in stars and the effects of Galactic evolution¹. These variations contrast with the uniformity of stable isotope abundances for many elements in the Solar System^{2,3}, which implies that processes efficiently homogenized dust and gas from different stellar sources within the young solar nebula. However, isotopic heterogeneity has been recognized on the subcentimetre scale in primitive meteorites^{4,5}, indicating that these preserve a compositional memory of their stellar sources. Small differences in the abundance of stable molybdenum isotopes in bulk rocks of some primitive^{6–8} and differentiated^{7,9} meteorites, relative to terrestrial Mo, suggest large-scale Mo isotopic heterogeneity between some inner Solar System bodies, which implies physical conditions that did not permit efficient mixing of gas and dust. Here we report Mo isotopic data for bulk samples of primitive and differentiated meteorites that show no resolvable deviations from terrestrial Mo. This suggests efficient mixing of gas and dust in the solar nebula at least to 3 AU from the Sun, possibly induced by magnetohydrodynamic instabilities¹⁰. These mixing processes must have occurred before isotopic fractionation of gas-phase elements and volatility-controlled chemical fractionations were established.

The preservation of heterogeneous isotopic compositions of variable nucleosynthetic origins in some components of primitive meteorites, such as pre-solar grains and some refractory condensates from the solar nebula^{4,5}, reflects the lack of equilibration of these components with gas and dust of average solar composition in cooler parts of the early solar accretion disk. Isotopic heterogeneities between larger Solar System bodies that do not reflect radiogenic ingrowth or mass-dependent isotopic fractionation have long been noted for elements that predominantly occur in the gas phase, such as O (ref. 11), C and N (ref. 12), and some noble gases⁴. These heterogeneities have been ascribed to late injection of isotopes synthesized in a nearby supernova, photochemical isotope fractionation during solid–gas or pure gas-phase reactions, or production by cosmic-ray bombardment in the early Solar System^{11,13–15}. Little evidence for large-scale isotopic heterogeneity of this type has been found in bulk samples of meteorites for elements that are largely condensed at high temperatures—for example, Cr (ref. 16) and Ti (ref. 17).

The enrichments of p- and r-process isotopes over s-process-dominated isotopes of Mo reported from bulk samples of the carbonaceous chondrite Allende^{6–8} and some iron and stony iron meteorites contrast with the absence of such anomalies in other groups of iron meteorites and enstatite chondrites^{7,9,18}. Isotopes produced by p-process (proton capture or photodissociation of nuclei) and r-process (rapid neutron capture) are believed to form in supernovae, a different environment from the red giants which are the main source of s-process isotopes, which form by slow neutron capture and subsequent β -decay¹. If such heterogeneities are real, there should also be collateral effects for isotopes of other elements produced by the same nucleosynthetic processes. No such effects have yet been found for Ru and Zr, the nearest neighbours of Mo with more than one stable isotope^{19–21}. This inconsistency and

the significance of the scale of isotopic heterogeneity for constraining evolution models of the early solar nebula have inspired the present work.

New data for the Mo isotopic composition of iron meteorites, chondrites and refractory inclusions from Allende were obtained by negative thermal ionization mass spectrometry (N-TIMS) (Table 1). Abundances of all Mo isotopes in three IIAB iron meteorites overlap with terrestrial Mo within ± 1 ϵ -unit (1 ϵ -unit represents 1 part in

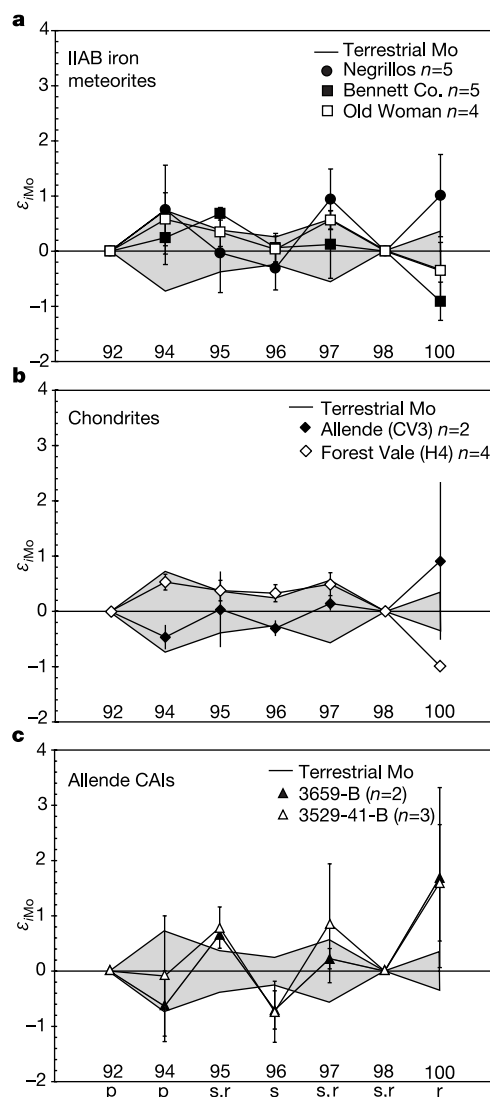


Figure 1 Isotopic composition of Mo in meteorites. Shown are deviations in ϵ -units (see Table 1) relative to terrestrial Mo ($\epsilon = 0$). **a**, IIAB iron meteorites. **b**, Chondrite bulk rocks. **c**, CAIs 3659-B and 3529-41-B from Allende, both studied previously for their Re–Os systematics²³. Mass discrimination was corrected using the exponential law²⁸ and a value of $^{92}\text{Mo}/^{98}\text{Mo} = 0.607898$, consistent with $^{94}\text{Mo}/^{98}\text{Mo} = 0.3802$ (ref. 27). The $^{92}\text{Mo}/^{98}\text{Mo}$ ratio was chosen for the mass spectrometer fractionation correction because it spans most of the large mass range of Mo. Only ^{100}Mo lies above this mass range. This may explain the larger scatter observed for samples at ^{100}Mo . Some previous studies have used $^{98}\text{Mo}/^{96}\text{Mo}$ for normalization. This normalization has the benefit of using two isotopes produced either partially or totally by s-process nucleosynthesis. It has the disadvantage that it spans less of the mass range of Mo, and could lead to inadequately corrected fractionation at the mass extremes. The grey field indicates $2\sigma_{\text{pop}}$ errors (2σ of the population, $n = 4$) for the standard. Data are means of multiple analyses with $2\sigma_{\text{pop}}$ uncertainties ($n =$ number of analyses, see Table 1). p, s and r indicate the nucleosynthetic processes that contribute to the different Mo isotopes (see text).

10,000 deviation of the sample from the terrestrial Mo standard; see Table 1 for exact definition). The one exception to this is Bennett County, which shows a slight negative deviation for ^{100}Mo compared to the Mo standard (Fig. 1a). The likely cause of this deviation is inadequate correction of mass spectrometer fractionation. Our results are consistent with the results of Yin *et al.*⁶, who reported no anomalies for two IIAB irons. They also found no anomalies for IA, IIIAB and IVB irons. These results, however, are in conflict with the results of Dauphas *et al.*⁷, who reported enrichments of ~ 0.5 to 2.5 ϵ -units for ^{92}Mo , ^{94}Mo , ^{95}Mo and ^{97}Mo for most of the IIAB irons they examined. They reported similar, albeit variable, enrichments for IIIAB, IIICD, IVA and IVB irons. Another group found large enrichments for ^{92}Mo and ^{94}Mo , relative to ^{96}Mo , for some IA and IVA iron meteorites⁹. True isotopic differences among different pieces of the same IIAB iron, or among different IIAB irons is very improbable because the IIAB group probably crystallized as a fractional crystallization sequence from a common, isotopically uniform liquid that comprised the metallic core of a small asteroid²². We conclude that the discrepancy between these studies is analytical. New data for the carbonaceous chondrite Allende (CV3) and the ordinary chondrite Forest Vale (H4) also overlap with

terrestrial Mo at the $\pm 1\epsilon$ level or better (Fig. 1b). Again, the exception is ^{100}Mo for Forest Vale, which like Bennett County shows a slight negative deviation of about 1ϵ unit.

In contrast to the results of both Yin *et al.*⁶ and Dauphas *et al.*^{7,8}, our result for bulk Allende shows no evidence for a significant enrichment in the p- and r-process isotopes, relative to pure s-process ^{96}Mo . A major concern for high-precision isotope ratio measurements is instrumental fractionation correction. The selection of two isotopes for the mass fractionation correction should have no effect on the detection of nonlinear isotopic anomalies, assuming that the correction algorithm used correctly accounts for fractionation in all isotopes, and that the masses are free of isobaric interferences. Inadequacies in the correction, however, could create anomalies that are the result of instrument bias. Because of this concern, we re-normalized some of our data to $^{98}\text{Mo}/^{96}\text{Mo}$, the ratio used in two previous studies^{6–8}. For the different normalizing ratio, our bulk Allende data show an Earth-like Mo isotopic composition (Fig. 2). Thus, selection of isotopes for fractionation correction probably is not the cause of the discrepancies.

The discrepancies between the present and previous results for the Allende bulk rock may either reflect the lithological heterogeneity of

Table 1 Mo isotopic composition of terrestrial standard and meteorites

Sample	$\epsilon^{92}\text{Mo}$	$\epsilon^{94}\text{Mo}$	$\epsilon^{96}\text{Mo}$	$\epsilon^{97}\text{Mo}$	$\epsilon^{100}\text{Mo}$	$^{101}\text{Ru}/^{95}\text{Mo}$
Terrestrial Mo standard:						
3-4/2002	0.00 $\pm$ 0.98	0.00 $\pm$ 0.79	0.00 $\pm$ 0.59	0.00 $\pm$ 0.78	0.00 $\pm$ 1.70	
7/2002	0.00 $\pm$ 0.86	0.00 $\pm$ 0.44	0.00 $\pm$ 0.29	0.00 $\pm$ 0.66	0.00 $\pm$ 0.40	
IIAB Iron meteorites:						
Negrillos	-0.18	0.25	-0.33	0.75	2.06	0.000072
Negrillos	0.55	-0.04	-0.35	0.31	0.57	0.000070
Negrillos	2.16	0.59	0.17	1.38	0.26	0.000067
Negrillos	0.21	-1.40	-1.00	1.76	1.75	0.000064
Negrillos	1.03	0.44	-0.03	0.52	0.40	0.000066
Mean ($n = 5$)	0.75 $\pm$ 0.81	-0.03 $\pm$ 0.72	-0.31 $\pm$ 0.40	0.94 $\pm$ 0.54	1.01 $\pm$ 0.74	0.000068 $\pm$ 3
Bennett County	0.12	0.54	0.09	-0.16	-1.39	0.000062
Bennett County	0.04	0.82	0.00	0.16	-1.03	0.000063
Bennett County	-0.40	0.62	-0.36	-0.76	-0.41	0.000044
Bennett County	1.08	0.81	0.44	1.12	-0.63	0.000064
Bennett County	0.37	0.61	0.15	0.23	-1.08	0.000064
Mean ($n = 5$)	0.24 $\pm$ 0.49	0.68 $\pm$ 0.11	0.06 $\pm$ 0.26	0.12 $\pm$ 0.61	-0.91 $\pm$ 0.35	0.000059 $\pm$ 8
Old Woman	0.76	-0.04	0.08	0.54	0.17	0.000050
Old Woman	1.16	0.47	0.28	0.82	-0.02	0.000066
Old Woman	0.07	0.44	0.09	0.46	-0.66	0.000049
Old Woman	0.33	0.51	-0.30	0.44	-0.91	0.000061
Mean ($n = 4$)	0.58 $\pm$ 0.48	0.34 $\pm$ 0.26	0.04 $\pm$ 0.24	0.56 $\pm$ 0.17	-0.35 $\pm$ 0.51	0.000056 $\pm$ 8
Chondrites:						
Allende (CV3)	-0.58	0.38	-0.23	0.08	0.20	0.000075
Allende (CV3)	-0.36	-0.31	-0.38	0.21	1.63	0.000068
Mean ($n = 2$)	-0.47 $\pm$ 0.22	0.04 $\pm$ 0.69	-0.30 $\pm$ 0.14	0.14 $\pm$ 0.13	0.91 $\pm$ 1.43	0.000072 $\pm$ 7
Forest Vale (H4)	0.46	0.41	0.19	0.27	-1.00	0.000072
Forest Vale (H4)	0.39	0.41	0.20	0.43	-1.06	0.000075
Forest Vale (H4)	0.54	0.12	0.46	0.77	-0.99	0.000062
Forest Vale (H4)	0.72	0.57	0.46	0.49	-0.95	0.000070
Mean ($n = 4$)	0.53 $\pm$ 0.14	0.38 $\pm$ 0.19	0.33 $\pm$ 0.15	0.49 $\pm$ 0.21	-1.00 $\pm$ 0.05	0.000070 $\pm$ 5
Allende CAIs:						
CAI 3529-41-B	-0.96	0.64	-0.88	0.13	0.88	0.000080
CAI 3529-41-B	-0.32	0.69	-0.53	0.32	2.51	0.000053
Mean ($n = 2$)	-0.64 $\pm$ 0.64	0.66 $\pm$ 0.05	-0.70 $\pm$ 0.35	0.23 $\pm$ 0.18	1.69 $\pm$ 1.63	0.000066 $\pm$ 27
CAI 3658-B	-0.26	0.42	-1.09	0.49	1.00	0.000083
CAI 3658-B	-0.93	0.90	-0.93	0.18	1.15	0.000074
CAI 3658-B	0.93	1.04	-0.19	1.92	2.65	0.000073
Mean ($n = 3$)	-0.09 $\pm$ 1.09	0.79 $\pm$ 0.37	-0.74 $\pm$ 0.55	0.86 $\pm$ 1.08	1.60 $\pm$ 1.05	0.000077 $\pm$ 6

Iron meteorites (1–3 g) were dissolved in concentrated HCl. Chondrites (~ 1 g) and CAIs were digested overnight or for 2 days (CAIs) at 235°C in borosilicate Carius tubes using concentrated hydrochloric and nitric acids in a 1:2 proportion. Molybdenum was separated from sample matrices by standard ion exchange procedures. Molybdenum was loaded on degassed Re filaments, metallized for 1 h, and covered with $\text{La}(\text{NO}_3)_3$. Molybdenum isotopic compositions were measured on the UMD Sector 54 multi collector TIMS in the form of MoO_3^- at $1,370$ – $1,420^\circ\text{C}$, and corrected for oxygen isotopic compositions using data from ref. 26. Signal intensities on mass 146 (^{96}Mo) were 200 – $1,000$ mV for Mo standards, and 100 – 900 mV for samples. Data were collected in two cycles. Molybdenum isotopes were all measured in the first cycle. Ruthenium interferences were monitored in the second cycle. Interference corrections on ^{96}Mo , ^{98}Mo and ^{100}Mo from Ru were small or insignificant. No interferences from Zr isotopes were detected. Total blanks were <10 ng, insignificant for the ~ 1 μg of Mo analysed for each aliquot processed. The Mo isotopic composition of a terrestrial ammonium molybdate solution from Fisher (700 – $1,000$ ng of Mo) was monitored as terrestrial reference material ('standard'). Our results for the Mo isotopic composition of the ammonium molybdate solution overlap within 1.5 ϵ -units with the data by Qi Lu and Masuda²⁷. Molybdenum isotopic data are reported as deviations in ϵ -units from the terrestrial Mo standard ($\epsilon = 0$), measured at the University of Maryland. $\epsilon = [(^{92}\text{Mo}/^{96}\text{Mo})_{\text{sa}}/ (^{92}\text{Mo}/^{96}\text{Mo})_{\text{std}} - 1] \times 10^4$ with $i = 92, 94, 95, 96, 97, 98, 100$, and subscripts sa and std indicating sample and terrestrial standard, respectively. n , Number of analyses used to calculate mean. Mean isotopic ratios and 2σ for Mo standard runs during the two run periods ($n = 9$, March–April 2002; $n = 6$, July 2002) are $^{94}\text{Mo}/^{96}\text{Mo} = 0.380206 \pm 75$ and 0.380175 ± 65 ; $^{95}\text{Mo}/^{96}\text{Mo} = 0.655942 \pm 104$ and 0.655972 ± 58 ; $^{96}\text{Mo}/^{96}\text{Mo} = 0.688133 \pm 81$ and 0.688115 ± 40 ; $^{97}\text{Mo}/^{96}\text{Mo} = 0.394971 \pm 62$ and 0.394997 ± 53 ; $^{100}\text{Mo}/^{96}\text{Mo} = 0.400252 \pm 136$ and 0.400201 ± 32 , respectively (uncertainties refer to the last digits of the ratio). Ruthenium interferences were $^{101}\text{Ru}/^{95}\text{Mo} = 0.000067 \pm 25$ and 0.000076 ± 19 , respectively. Because samples were analysed 2–5 times, the analytical precision for samples and standards are given as $2\sigma_{\text{pop}}$, calculated with $n = 2$ –5 for samples and $n = 4$ for the standards. Overall agreement between $2\sigma_{\text{pop}}$ of standards and samples supports this approach. Negrillos data were normalized to the mean of standard runs during the period March–April 2002. Other meteorite data were normalized to the mean of standard runs during the period July 2002.

this meteorite, or, alternatively, selective digestion of components, or both. Lithological heterogeneity would suggest that the component showing the p- and r-isotope enrichment ('Mo-w'⁶⁻⁸) and its s-enriched complement ('Mo-m'⁸) are associated with major lithological components of Allende whose abundances vary among different pieces of the meteorite. Selective digestion of certain Mo-rich components could also account for the discrepancies. Acid digestion in Teflon beakers at relatively low temperatures, as used by the previous studies, may not access all Mo. This could lead to a selective accessing of Mo from a component with highly anomalous Mo. Complete digestion of Allende is difficult using this digestion technique because it contains abundant refractory inclusions (calcium–aluminium-rich inclusions, CAIs) with refractory oxides, such as spinel, that are difficult to digest. In this context it is noteworthy that a bulk sample of the carbonaceous chondrite Orgueil (CI), which does not contain CAIs, did not show nucleosynthetic anomalies, although its components have been shown to be isotopically heterogeneous⁸. In the present study chondrites were digested at elevated temperatures and pressures in sealed Carius tubes, a technique that has been shown to access >90% of the most refractory elements in CAIs²³. Thus, our results may be more representative of the bulk chondrites. This interpretation is also consistent with the observation that enstatite chondrites¹⁸ and metal from an ordinary chondrite⁶ also show no deviations from an Earth-like Mo isotopic composition.

New Mo isotopic data for two CAIs from Allende overlap with terrestrial Mo within ± 1 ϵ -unit (Fig. 1c). However, we note that, if compared to the s-process isotope ⁹⁶Mo, the CAIs show enhancements of ~ 1.5 ϵ in ⁹⁵Mo, ~ 0.9 to 1.6 ϵ in ⁹⁷Mo, and ~ 1.6 ϵ in ¹⁰⁰Mo. These enhancements persist, regardless of which normalizing ratio is used. Yin *et al.*⁶ found a very similar pattern for Mo isotopes in a CAI from Allende, and interpreted it to reflect the presence of r-process components in the CAI precursor. Nucleosynthetic anomalies in some CAIs are well documented for Ti, Ca, Cr, Ni, Zn, Zr, Ba and other elements⁵. The Mo isotopic results obtained on chondrite components, as manifested in data on leachates and residues from Allende and Orgueil⁸ and CAIs from Allende, indicate the presence of at least three prevalent components in some carbonaceous chondrites, Mo-w, Mo-m, and a component in some CAIs enriched in r-process isotopes. The latter observation appears to be consistent with enrichment of r-process isotopes of Ba and Zr in some other CAIs^{5,24}. This and anomalies for some isotopes below the iron abundance peak⁵ suggest that the precursor materials of these CAIs may have contained a larger

fraction of stardust from supernovae than is present in the Solar System on average. The consistent recovery of the Mo-w pattern in leaching experiments⁸ may indicate that the Mo-w signature belongs to a component in which p- and r-process signatures are inseparably linked. These observations are reminiscent of the inseparable, p- and r-process signature for a noble gas fraction ('Xe-HL') in pre-solar SiC and diamond residues from carbonaceous chondrites⁴. However, it has been argued on the basis of mass balance that such grains cannot account for the Mo-w signature^{6,7}. On the other hand, leaching experiments⁸ did not identify the Mo r-process pattern shown by the CAIs, illustrating that more than one approach is necessary to unequivocally separate the different components.

The new Mo isotopic data provide no evidence for large-scale and systematic isotopic heterogeneities in inner Solar System bodies at the ϵ level. Homogeneous Ru and Zr isotopic compositions of bulk meteorites¹⁹⁻²¹ are consistent with the new Mo isotopic data. Our data support models that suggest effective mixing and homogenization of gas and dust, at a relatively early stage of the evolution of the solar nebula. Homogenization in the inner Solar System may have been accomplished by turbulent motion caused by magnetic instabilities within the proto-solar accretion disk¹⁰. However, these instabilities entail a certain amount of ionization of the gas, and it is unknown if they occurred as far out as the location of the asteroid belt. Preservation of isotopically homogenized material that later constituted inner Solar System bodies requires conditions and grain sizes that prevent these materials from accretion into the protosun. How this was accomplished is uncertain. It has been suggested that some chondrite components, such as CAIs and chondrules, underwent high-temperature processing close to the protosun, and were then swept out into the 3 AU region by powerful winds, where they mixed with other chondrite components²⁵. Such a scenario would result in arbitrary mixtures of the isotopically heterogeneous components, and is difficult to reconcile with the homogeneous Mo isotopic compositions of Earth and bodies from the asteroid belt.

Homogeneous Mo isotopic compositions in bulk rocks of chondrites suggest that the various chondrite components were processed in a common environment. Because chondrites such as Allende and Forest Vale have homogeneous Mo isotopic composition, but heterogeneous O isotopic compositions, it appears most plausible that the homogeneous isotopic compositions of elements that condense at high temperatures were established before the creation of isotopic heterogeneity of elements that occur predominantly in the gas phase. Future studies need to explore further the physical conditions that facilitate mixing and subsequent preservation of gas and dust in the solar nebula, assess mixing and homogenization in molecular cloud cores before protostellar collapse, and resolve the origin of heterogeneities in ⁵⁴Cr and ⁵⁰Ti in bulk rocks of some primitive chondrites. □

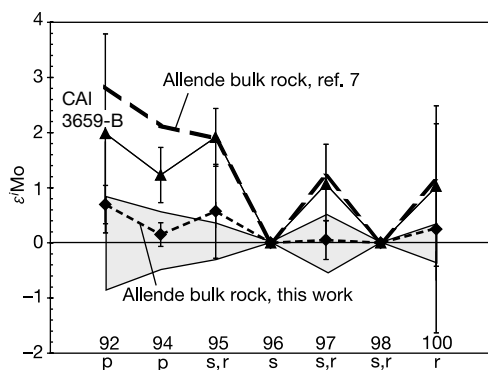


Figure 2 Isotopic composition of Mo in bulk Allende and CAI 3659-B. The data are displayed as deviations from terrestrial Mo ($\epsilon = 0$), using ⁹⁸Mo/⁹⁶Mo = 1.453173 (ref. 27) to correct for mass discrimination. Shown for comparison is the mean of three Allende bulk rock analyses from Dauphas *et al.*⁷. Note the Earth-like isotopic composition of the new Allende bulk rock data.

Received 27 January; accepted 5 August 2003; doi:10.1038/nature01975.

1. Wallerstein, G. *et al.* Synthesis of the elements in stars: forty years of progress. *Rev. Mod. Phys.* **69**, 995–1084 (1997).
2. Suess, H. E. & Urey, H. C. Abundances of the elements. *Rev. Mod. Phys.* **28**, 53–74 (1956).
3. Anders, E. & Grevesse, N. Abundances of the elements: Meteoritic and solar. *Geochim. Cosmochim. Acta* **53**, 197–214 (1989).
4. Anders, E. & Zinner, E. Interstellar grains in primitive meteorites: Diamond, silicon carbide, and graphite. *Meteoritics* **28**, 490–514 (1993).
5. Harper, C. L. in *Nuclei in the Cosmos* (eds Käppeler, F. & Wisshak, K.) 113–126 (IOP Publishing, Bristol, 1993).
6. Yin, Q., Jacobsen, S. B. & Yamashita, K. Diverse supernova sources of pre-solar material inferred from molybdenum isotopes in meteorites. *Nature* **415**, 881–883 (2002).
7. Dauphas, N., Marty, B. & Reisberg, L. Molybdenum evidence for inherited planetary scale isotope heterogeneity of the protosolar nebula. *Astrophys. J.* **565**, 640–644 (2002).
8. Dauphas, N., Marty, B. & Reisberg, L. Molybdenum nucleosynthetic dichotomy revealed in primitive meteorites. *Astrophys. J.* **569**, L139–L142 (2002).
9. Qi Lu & Masuda, A. in *Origin of Elements in the Solar System* (ed. Manuel, O.) 385–400 (Kluwer/Plenum, New York, 2000).
10. Stone, J. M., Gammie, C. F., Balbus, S. A. & Hawley, J. F. in *Protostars and Planets IV* (eds Mannings, V., Boss, A. P. & Russell, S. S.) 589–611 (Univ. Arizona Press, Tucson, 2000).

11. Clayton, R. N. Oxygen isotopes in meteorites. *Annu. Rev. Earth Planet. Sci.* **21**, 115–149 (1993).
12. Lewis, R. S., Anders, E., Wright, I. P., Norris, S. J. & Pillinger, C. T. Isotopically anomalous nitrogen in primitive meteorites. *Nature* **305**, 767–771 (1983).
13. Thieme, M. H. & Heidenreich, J. E. III The mass-independent fractionation of oxygen: A novel isotope effect and its possible cosmochemical implications. *Science* **219**, 1073–1075 (1983).
14. Clayton, R. N. Self-shielding in the solar nebula. *Nature* **415**, 860–861 (2002).
15. Lee, T., Shu, F. H., Shang, H., Glassgold, A. E. & Rehm, K. E. Protostellar cosmic rays and extinct radioactivities in meteorites. *Astrophys. J.* **506**, 898–912 (1998).
16. Shukolyukov, A. & Lugmair, G. W. Manganese-chromium isotope systematics of Ivuna, Kainsaz and other carbonaceous chondrites. *Lunar Planet. Sci.* **XXXIV**, 1279 (2003).
17. Niederer, F. R., Papanastassiou, D. A. & Wasserburg, G. J. Absolute isotopic abundances of Ti in meteorites. *Geochim. Cosmochim. Acta* **49**, 835–851 (1985).
18. Dauphas, N., Marty, B. & Reisberg, L. Inference on terrestrial genesis from molybdenum isotope systematics. *Geophys. Res. Lett.* **29**, 1084 (2002).
19. Becker, H. & Walker, R. J. Ruthenium isotopic composition of terrestrial materials, iron meteorites and chondrites. *Lunar Planet. Sci.* **XXXIII**, 1018 (2002).
20. Becker, H. & Walker, R. J. In search of extant Tc in the early solar system: ^{98}Ru and ^{99}Ru abundances in iron meteorites and chondrites. *Chem. Geol.* **196**, 43–56 (2003).
21. Schönbachler, M., Lee, D.-C., Halliday, A. N. & Rehkämper, M. Uniformity of zirconium isotopic compositions in the inner solar system. *Lunar Planet. Sci.* **XXXIII**, 1283 (2002).
22. Pernicka, E. & Wasserburg, G. J. Ru, Os, Pt and Au in iron meteorites. *Geochim. Cosmochim. Acta* **51**, 1717–1726 (1987).
23. Becker, H., Morgan, J. W., Walker, R. J., MacPherson, G. J. & Grossman, J. N. Rhenium-osmium systematics of calcium-aluminium-rich inclusions in carbonaceous chondrites. *Geochim. Cosmochim. Acta* **65**, 3379–3390 (2001).
24. McCulloch, M. T. & Wasserburg, G. J. Barium and neodymium isotopic anomalies in the Allende meteorite. *Astrophys. J.* **220**, L15–L19 (1978).
25. Shu, F. H., Shang, H. & Lee, T. Toward an astrophysical theory of chondrites. *Science* **271**, 1545–1552 (1996).
26. Nier, A. O. A redetermination of the relative abundances of the isotopes of carbon, nitrogen, oxygen, argon and potassium. *Phys. Rev.* **77**, 789–793 (1950).
27. Qi Lu & Masuda, A. The isotopic composition and atomic weight of molybdenum. *Int. J. Mass Spectrom. Ion Proc.* **130**, 65–72 (1994).
28. Russell, W. A., Papanastassiou, D. A. & Tombrello, T. A. Ca isotope fractionation on the Earth and other solar system materials. *Geochim. Cosmochim. Acta* **42**, 1075–1090 (1978).

Acknowledgements We thank R. Clarke and G. MacPherson for providing the meteorite samples, N. Dauphas, H. Palme, D. Papanastassiou, A. Pietruszka and Q. Yin for discussions, and P. Tomascak for comments on the manuscript and technical support in the laboratory. Comments by S. Jacobsen have improved the presentation of various issues in this manuscript.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.B. (hbecker@geol.umd.edu).

Quantum dynamics of a single vortex

A. Wallraff*, A. Lukashenko, J. Lisenfeld, A. Kemp, M. V. Fistul, Y. Koval & A. V. Ustinov

Physikalisches Institut III, Universität Erlangen-Nürnberg, D-91058 Erlangen, Germany

* Present address: Department of Applied Physics, Yale University, New Haven, Connecticut 06520, USA

Vortices occur naturally in a wide range of gases and fluids, from macroscopic to microscopic scales. In Bose–Einstein condensates of dilute atomic gases¹, superfluid helium² and superconductors, the existence of vortices is a consequence of the quantum nature of the system. Quantized vortices of supercurrent³ are generated by magnetic flux penetrating the material, and play a key role in determining the material properties⁴ and the performance of superconductor-based devices^{5,6}. At high temperatures the dynamics of such vortices are essentially classical, while at low temperatures previous experiments have suggested collective quantum dynamics^{7,8}. However, the question of whether vortex tunnelling occurs at low temperatures has been addressed only for large collections of vortices. Here we study the quantum dynamics of an individual vortex in a superconducting Josephson junction. By measuring the statistics of the vortex escape from a controllable pinning potential, we demonstrate the existence of

quantized levels of the vortex energy within the trapping potential well and quantum tunnelling of the vortex through the pinning barrier.

The object that we have studied in our experiments is a vortex of electric current with a spatial extent of several tens of micrometres formed in a long superconducting tunnel junction. The electro-dynamics of such a Josephson junction is governed by the phase difference ϕ between the macroscopic wavefunctions describing the superconducting condensate in the two electrodes⁹. It is a well established fact that the variable ϕ does display macroscopic quantum properties in point-like junctions at very low temperatures^{10,11}. Such macroscopic quantum phenomena are currently exploited for quantum information processing¹² using superconducting devices^{10,13–17}. In extended one- or two-dimensional Josephson junction systems, quantum tunnelling in real space is to be expected for superconducting vortices, which are particle-like collective excitations of the phase difference ϕ . Collective nonlinear excitations such as the vortex considered here are ubiquitous in solid state systems (for example, domain walls), biological systems (such as waves on membranes), and have even been considered as model systems in particle physics. The small value of the expected mass of the vortex studied here suggests that quantum effects are likely to occur with vortices at low temperatures. Dissipative vortex tunnelling has been put forward as a reason for the non-vanishing relaxation rate of the magnetization in type-II superconductors as the temperature T is lowered towards zero. However, the quantum vortex creep model⁸, which was suggested as an explanation for this behaviour, faced orders-of-magnitude discrepancies with experimental data¹⁸. Alternative classical explanations have been suggested more recently to explain the low-temperature relaxation¹⁹. Nonetheless, macroscopic quantum tunnelling was observed for states with many vortices in discrete arrays of small Josephson junctions^{20,21}. For typical arrays, the calculated vortex mass is about 500 times smaller than the electron mass⁷. All previous research in this area has focused on the collective behaviour of a large number of vortices. Until now there had been no direct experimental observation of tunnelling events of individual vortices.

Among many different vortex structures in superconductors, there is a rather special type of vortex in long Josephson junctions. These vortices have the characteristics of solitons—nonlinear waves that preserve their shape with time and propagate as ballistic particles²². These vortices are distinct from Abrikosov vortices in type-II superconductors, as they have no normal core and thus move with very low damping. In contrast to vortices in Josephson arrays⁷, solitons in uniform long junctions do not generate any radiation during their motion, and are well decoupled from other electromagnetic excitations in these systems. The quantum tunnelling of Josephson vortices in long junctions has been predicted theoretically^{23,24}. Here we present experimental observations of this effect.

We probe the quantum properties of a single Josephson vortex in a current-biased annular junction subject to an in-plane magnetic field H (Fig. 1c inset). The junction (of diameter $d = 100\ \mu\text{m}$ and width $w = 0.5\ \mu\text{m}$) is etched from a sputtered Nb/AlO_x/Nb thin-film trilayer which is patterned using electron-beam lithography²⁵. A photograph of the sample taken using an optical microscope is shown in Fig. 1a. Initially the vortex is topologically trapped in the junction by cooling the sample in a small perpendicular magnetic field H_{tr} which is generated using a separate pair of coils, the axis of which is perpendicular to the junction plane (Fig. 1b). A single vortex in an annular junction subject to an in-plane field behaves as a particle²⁶ in a tilted washboard potential^{27,28}. The component of the potential periodic in the vortex coordinate θ is due to the interaction $\mu \cdot H \propto \cos\theta$ of the vortex magnetic moment μ with the external magnetic field H (Fig. 1c). The tilt of the potential is proportional to the Lorentz force acting on the vortex which is induced by the bias current I applied to the junction. The vortex

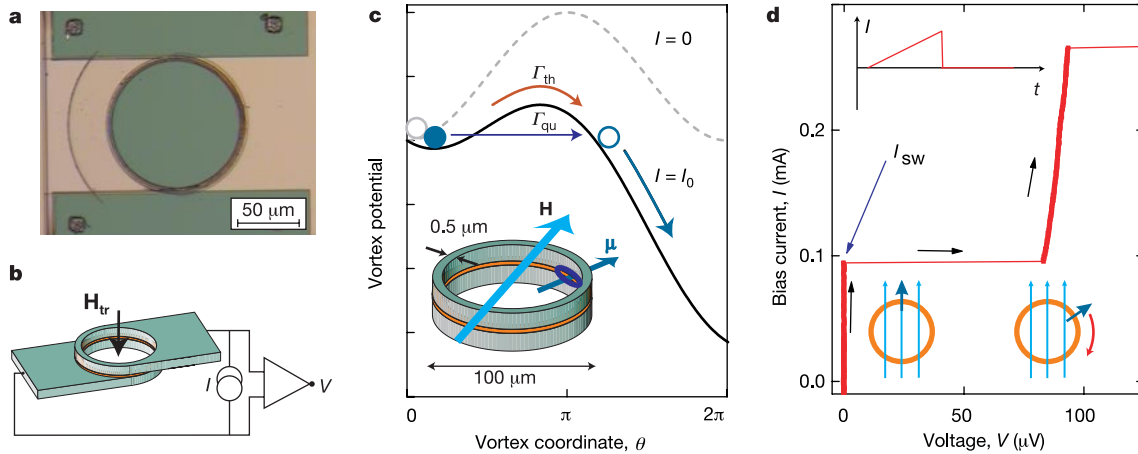


Figure 1 Sample, vortex potential and switching current measurement. **a**, Photograph of the sample taken with an optical microscope. **b**, Sketch of bias lead configuration and direction of trapping field H_{tr} . We use a bias lead geometry which minimizes self field of the junction but perpendicular to the bias leads. In this case, the vortex is depinned in a location along the junction where self fields are minimal. **c**, A vortex with magnetic moment μ trapped in an annular Josephson junction subject to an in-plane external magnetic field H . The junction width and diameter are indicated (see inset). The resulting vortex potential at zero bias current I (dashed line) and at finite bias I_0 (solid line) is plotted

versus the vortex coordinate θ in the annulus. Thermal and quantum escape of the vortex at rates Γ_{th} and Γ_{qu} , respectively are indicated. **d**, Current–voltage characteristic showing the vortex depinning from the field induced potential at a random value of bias I_{sw} when ramping up the bias current at a constant rate in a saw-tooth pattern (see upper inset). The transition of a pinned vortex state to a running vortex state is associated with a voltage appearing across the junction which is proportional to the vortex velocity (see lower insets). In experiment, the bias current is switched off immediately after a voltage is detected.

may escape from a well in the tilted potential by a thermally activated process or by quantum tunnelling. At low temperatures thermal activation is exponentially suppressed, and the escape occurs by quantum tunnelling. This process is identified by measuring the temperature dependence of the distribution P of depinning currents I_{sw} (refs 29, 30) of the vortex trapped in the junction. The $P(I_{sw})$ distributions are recorded by repeatedly ramping up the bias current at a constant rate $\dot{I}$ and recording the statistics of the current I_{sw} at which the vortex escapes from the well, which is associated with a switching of the junction from its zero-voltage state to a finite-voltage state (Fig. 1d). Our measurement technique and set-up have been tested and calibrated^{30,31} in experiments on macroscopic quantum tunnelling of the phase in small Josephson junctions¹¹.

The bias current I_p at which vortex tunnelling occurs with the highest probability (for a given bias current ramp rate) is found—as expected^{27,28}—to be proportional to the applied magnetic field (see below). This indicates that the shape of the potential is controlled in our experiment by both field and bias current. The effect of the bias-current-induced self magnetic field on the vortex potential has been minimized by using an appropriate bias lead configuration and field direction (Fig. 1b).

To search for quantum tunnelling of the vortex, the temperature dependence of the switching current distribution $P(I_{sw})$ has been measured. In Fig. 2a such distributions are shown for a magnetic field of $H = 0.9$ Oe applied in the plane of the junction. It is clearly observed that the distribution width σ decreases with temperature and then saturates at low T . In Fig. 2b, σ is plotted versus temperature on a double logarithmic scale for two different values of field. At high temperatures, the distribution width is temperature dependent, indicating the thermally activated escape of the vortex from the well. In the high temperature limit σ is, to a good approximation, proportional to $T^{2/3}$ as expected for a thermally activated escape of a particle from a washboard potential close to critical bias. The distribution width σ saturates at a temperature of about 100 mK. This behaviour indicates the cross-over of the vortex escape process from thermal activation to quantum tunnelling. At

temperatures below 100 mK, σ is constant and the escape is dominated by quantum tunnelling. As expected, the cross-over temperature T^* is dependent on magnetic field, but only rather weakly. We attribute this observation to the fact that the vortex may change its shape when traversing the barrier during the escape process. This aspect can not be captured in the single-particle model^{26,27}, but rather is a consequence of the fact that the vortex is a collective excitation.

To probe the energy levels of the vortex in the potential well, we have measured the vortex escape in the presence of microwave radiation, using spectroscopic techniques which we have extensively tested on small Josephson junctions³¹. At low temperatures and in the absence of microwave radiation, the vortex tunnels out of the ground state of the potential well into the continuum. The occupation of excited states is exponentially small, if the level separation is larger than the temperature. By irradiating the sample with microwaves, the vortex can be excited resonantly from the ground state to the first excited state (Fig. 3b inset). In this case, we observe a double peak structure in the switching current distribution (Fig. 3a). The peak at higher bias current is due to the tunnelling of the vortex from the ground state, whereas the peak at lower bias current corresponds to the tunnelling out of the first excited state.

The energy level separation scales both with bias current and applied magnetic field. To investigate this property, we have spectroscopically determined the separation between the ground and the first excited state by varying the microwave frequency and the magnetic field. For each value of the magnetic field, we have determined the resonance current I_r (see Fig. 3a) for a few different microwave frequencies ν . In Fig. 3b, the applied microwave frequency ν is plotted versus the resonance current I_r normalized by the depinning current $I_c(H)$ at that field in absence of microwaves and fluctuations. It is observed that all data points show the characteristic scaling of the energy level separation $\Delta E_{01}/h$ with the bias current as $\nu_{01}(H, I) = \nu_{01}(H, I = 0)(1 - (I/I_c(H))^2)^{1/4}$, as expected for the tilted washboard potential. The data at each field are fitted to this dependence using the characteristic frequency $\nu_{01}(H, I = 0)$ of vortex oscillations at zero current and the depin-

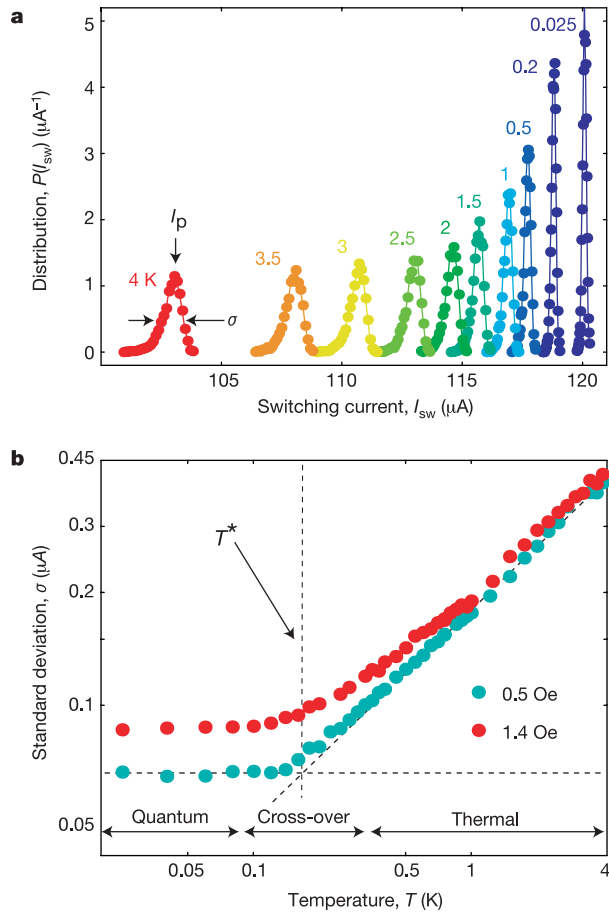


Figure 2 Thermal activation and quantum tunnelling. **a**, Switching current distributions $P(I_{sw})$ at magnetic field $H = 0.9$ Oe for bath temperatures T between 4.0 K and 25 mK. **b**, Standard deviation σ of $P(I_{sw})$ versus T for two values of field, indicating the cross-over in the vortex escape process from thermal activation to quantum tunnelling. The cross-over temperature range around T^* is indicated. We have carefully verified experimentally that the saturation of the distribution width with temperature is not induced by excess noise or heating of the sample. The resolution of our measurement set-up³⁰ is factor of 3 to 4 larger than the most narrow distribution widths measured in these experiments.

ning current $I_c(H)$ as fitting parameters, see dashed lines in Fig. 3b. The resonance frequency $\nu_{01}(H, I = 0)$ and the depinning current $I_c(H)$ in absence of fluctuations have been determined from the fit. As expected, the energy level separation ΔE_{01} increases with field and the depinning current I_c as determined from the spectroscopic data is linear in the field, see Fig. 3c, which is in excellent agreement with the current I_p measured directly in absence of microwaves.

From the resonance frequency at a given bias current we can estimate the cross-over temperature, which in the limit of small damping is given by $T^* \approx \hbar \nu_{01}(H, I)/2\pi k_B$. Thus we can compare the cross-over temperature extracted from the temperature dependence of the switching current distributions to the predictions based on the data extracted from spectroscopic measurements. For $\nu_{01}(H, I)$ between 10 and 13 GHz we find a value of T^* between approximately 75 and 100 mK, which is consistent with the measured saturation temperature in Fig. 2.

The quantum dynamics of vortices are important for evaluating the properties of superconducting materials and devices at low temperatures. Similar effects to those observed here in a superconductor may well be observable in atomic Bose–Einstein condensates or in superfluid helium. The results of our experiments may directly

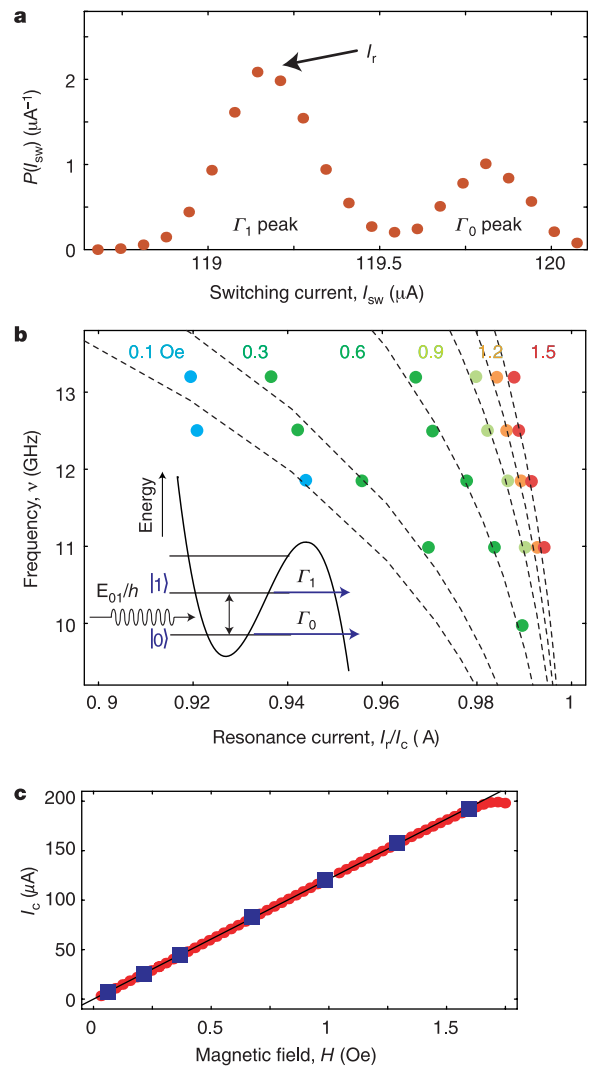


Figure 3 Vortex energy levels. **a**, $P(I_{sw})$ distribution at $H = 0.6$ Oe in the presence of microwave radiation of $\nu = 11$ GHz at $T = 25$ mK. The resonance current I_r at which tunnelling from the first excited state is most probable is indicated. **b**, Microwave frequency ν versus normalized resonance current $I_r/I_c(H)$ for magnetic fields H between 0.1 and 1.5 Oe. Dashed curves are fits to $\nu_{01}(H, I = 0)(1 - (I/I_c(H))^2)^{1/4}$. In the inset is indicated the tunnelling from the ground state at rate Γ_0 and from the first excited state populated by resonant microwave radiation at rate Γ_1 . **c**, Critical current I_c extracted from microwave spectroscopy (filled squares), most probable switching current I_p at $T = 25$ mK in absence of microwaves (solid circles) and fit (solid line) versus magnetic field H .

open a path towards quantum computation¹² using Josephson vortices in long junctions as qubits^{32,33}.

Received 31 March; accepted 17 June 2003; doi:10.1038/nature01826.

1. Abo-Shaeer, J. R., Raman, C., Vogels, J. M. & Ketterle, W. Observation of vortex lattices in Bose–Einstein condensates. *Science* **292**, 476–479 (2001).
2. Blaauwgeers, R. *et al.* Double-quantum vortex in superfluid ³He-A. *Nature* **404**, 471–473 (2000).
3. Huebener, R. P. *Magnetic Flux Structures in Superconductors* (Springer, Berlin, 2001).
4. Bugoslavsky, Y., Perkins, G. K., Qi, X., Cohen, L. F. & Caplin, A. D. Vortex dynamics in superconducting MgB₂ and prospects for applications. *Nature* **410**, 563–565 (2001).
5. Nori, F. & Savel'ev, S. Experimentally realizable devices for controlling the motion of magnetic flux quanta in anisotropic superconductors. *Nature Mater.* **1**, 179–184 (2002).
6. Lee, C.-S., Janko, B., Derenyi, I. & Barabesi, A.-L. Reducing the vortex density in superconductors using the ratchet effect. *Nature* **400**, 337–340 (1999).
7. Fazio, R. & van der Zant, H. J. S. Quantum phase transitions and vortex dynamics in superconducting networks. *Phys. Rep.* **355**, 235–334 (2001).
8. Blatter, G., Feigel'man, M. V., Geshkenbein, A. I. & Vinokur, V. M. Vortices in high-temperature superconductors. *Rev. Mod. Phys.* **66**, 1125–1388 (1994).
9. Tinkham, M. *Introduction to Superconductivity* (McGraw-Hill International Editions, New York, 1996).

10. Clarke, J., Cleland, A. N., Devoret, M. H., Esteve, D. & Martinis, J. M. Quantum mechanics of a macroscopic variable: The phase difference of a Josephson junction. *Science* **239**, 992–997 (1988).
11. Devoret, M. H. *et al.* in *Quantum Tunneling in Condensed Media*, (eds Kagan, Yu. & Leggett, A. J.) (North-Holland, Amsterdam, 1992).
12. Bennett, C. H. & DiVincenzo, D. P. Quantum information and computation. *Nature* **404**, 247–255 (2000).
13. Nakamura, Y., Pashkin, Y. A. & Tsai, J. S. Coherent control of macroscopic quantum states in a single-Cooper-pair box. *Nature* **398**, 786–788 (1999).
14. Friedman, J. R., Patel, V., Chen, W., Tolpygo, S. K. & Lukens, J. E. Quantum superposition of distinct macroscopic states. *Nature* **406**, 43–46 (2000).
15. van der Wal, C. H. *et al.* Quantum superposition of macroscopic persistent-current states. *Science* **290**, 773–777 (2000).
16. Pashkin, Y. A. *et al.* Quantum oscillations in two coupled charge qubits. *Nature* **421**, 823–826 (2003).
17. Chiorescu, I., Nakamura, Y., Harmans, C. J. P. M. & Mooij, J. E. Coherent quantum dynamics of a superconducting flux qubit. *Science* **299**, 1869–1871 (2003).
18. Hoekstra, H. F. T. *et al.* General features of quantum creep in high- T_c superconductors. *Phys. Rev. Lett.* **80**, 4293–4296 (1998).
19. Nicodemi, M. & Jensen, H. J. Creep of superconducting vortices in the limit of vanishing temperature: A fingerprint of off-equilibrium dynamics. *Phys. Rev. Lett.* **86**, 4378–4381 (2001).
20. van der Zant, H. S. J., Fritschy, J. F. C., Orlando, T. P. & Mooij, J. E. Dynamics of vortices in underdamped Josephson-junction arrays. *Phys. Rev. Lett.* **66**, 531–2534 (1991).
21. Tighe, T. S., Johnson, A. T. & Tinkham, M. Vortex motion in two-dimensional arrays of small, underdamped Josephson junctions. *Phys. Rev. B* **44**, 10286–10290 (1991).
22. Ustinov, A. V. Solitons in Josephson junctions. *Physica D* **123**, 315–329 (1998).
23. Kato, T. & Imada, M. Macroscopic quantum tunneling of a fluxon in a long Josephson junction. *J. Phys. Soc. Jpn* **65**, 2963–2975 (1996).
24. Shnirman, A., Ben-Jacob, E. & Malomed, B. A. Tunneling and resonant tunneling of fluxons in a long Josephson junction. *Phys. Rev. B* **56**, 14677–14685 (1997).
25. Koval, Y. *et al.* Narrow long Josephson junctions. *IEEE Trans. Appl. Supercond.* **9**, 3957–3961 (1999).
26. McLaughlin, D. W. & Scott, A. C. Perturbation analysis of fluxon dynamics. *Phys. Rev. A* **18**, 1652–1980 (1978).
27. Grönbech-Jensen, N., Lomdahl, P. & Samuelsen, M. Phase locking of long annular Josephson junctions coupled to an external rf magnetic field. *Phys. Lett. A* **154**, 14–18 (1991).
28. Ustinov, A. V. & Thyssen, N. Experimental study of fluxon dynamics in a harmonic potential well. *J. Low Temp. Phys.* **106**, 193–200 (1997).
29. Fulton, T. A. & Dunkleberger, L. N. Lifetime of the zero-voltage state in Josephson tunnel junctions. *Phys. Rev. B* **9**, 4760–4768 (1974).
30. Wallraff, A. *et al.* Switching current measurements of large area Josephson tunnel junctions. *Rev. Sci. Instrum.* **74**, 3740–3748 (2003).
31. Wallraff, A., Duty, T., Lukashenko, A. & Ustinov, A. V. Multi-photon transitions between energy levels in a current-biased Josephson tunnel junction. *Phys. Rev. Lett.* **90**, 037003 (2003).
32. Wallraff, A., Koval, Y., Levitchev, M., Fistul, M. V. & Ustinov, A. V. Annular long Josephson junctions in a magnetic field: Engineering and probing the fluxon potential. *J. Low Temp. Phys.* **118**, 543–553 (2000).
33. Kemp, A., Wallraff, A. & Ustinov, A. V. Josephson vortex qubit: Design, preparation and read-out. *Phys. Status Solidi B* **233**, 472–481 (2002).
34. Martucciello, N. *et al.* Fluxon dynamics in long annular Josephson tunnel junctions. *Phys. Rev. B* **57**, 5444–5449 (1998).

Acknowledgements We thank G. Blatter, T. Kato, G. Schön and S. Shnirman for discussions. This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.W. (andreas.wallraff@yale.edu).

Efficient bulk heterojunction photovoltaic cells using small-molecular-weight organic thin films

Peter Peumans, Soichi Uchida & Stephen R. Forrest

Center for Photonics and Optoelectronic Materials (POEM), Department of Electrical Engineering and the Princeton Materials Institute, Princeton University, Princeton, New Jersey 08544, USA

The power conversion efficiency of small-molecular-weight and polymer organic photovoltaic cells has increased steadily over the past decade. This progress is chiefly attributable to the introduction of the donor–acceptor heterojunction^{1,2} that functions as a dissociation site for the strongly bound photogenerated excitons. Further progress was realized in polymer devices through use of

blends of the donor and acceptor materials^{3–5}: phase separation during spin-coating leads to a bulk heterojunction that removes the exciton diffusion bottleneck by creating an interpenetrating network of the donor and acceptor materials. The realization of bulk heterojunctions using mixtures of vacuum-deposited small-molecular-weight materials has, on the other hand, posed elusive: phase separation induced by elevating the substrate temperature inevitably leads to a significant roughening of the film surface and to short-circuited devices. Here, we demonstrate that the use of a metal cap to confine the organic materials during annealing prevents the formation of a rough surface morphology while allowing for the formation of an interpenetrating donor–acceptor network. This method results in a power conversion efficiency 50 per cent higher than the best values reported for comparable bilayer devices, suggesting that this strained annealing process could allow for the formation of low-cost and high-efficiency thin film organic solar cells based on vacuum-deposited small-molecular-weight organic materials.

The external quantum efficiency of a photovoltaic cell based on exciton dissociation at a donor–acceptor interface is⁶ $\eta_{\text{EQE}} = \eta_{\text{A}} \times \eta_{\text{ED}} \times \eta_{\text{CC}}$. Here, η_{A} is the absorption efficiency. The exciton diffusion efficiency, η_{ED} , is the fraction of photogenerated excitons that reaches a donor–acceptor interface before recombining. The carrier collection efficiency, η_{CC} , is the probability that a free carrier, generated at a donor–acceptor interface by dissociation of an exciton, reaches its corresponding electrode. Typically, in bilayer donor–acceptor photovoltaic cells with a total thickness, L , of the order of the optical absorption length, L_{A} , we have $\eta_{\text{A}} = 1 - \exp(-L/L_{\text{A}}) > 50\%$ if optical interference effects are ignored, and $\eta_{\text{CC}} \approx 100\%$. However, since the exciton diffusion length (L_{D}) is typically an order of magnitude smaller than L_{A} , a large fraction of the photogenerated excitons remains unused for photocurrent generation⁶ (Fig. 1a), limiting η_{EQE} and hence the power conversion efficiency, η_{P} , of this type of planar junction cell.

In polymer photovoltaic cells, the exciton diffusion bottleneck has been removed through the introduction of bulk heterojunctions^{4,5} (Fig. 1b). In a bulk heterojunction, the donor–acceptor interface is highly folded such that photogenerated excitons find an interface within a distance L_{D} of their generation site. Currently, state-of-the-art bulk heterojunction polymer photovoltaic cells have power conversion efficiencies of up to 3.5% (refs 7–9). The bulk heterojunction is fabricated by spin-coating a mixture of the donor and acceptor materials. During spin coating and solvent evaporation, the donor and acceptor materials phase separate, creating an intricate interpenetrating network.

Attempts to achieve a bulk heterojunction through codeposition of the donor and acceptor materials yielded devices with η_{P} values falling short of those achievable in optimized bilayer cells using the same materials^{10–16}. Strong quenching of the photoluminescence in mixed materials indicates that $\eta_{\text{ED}} \approx 100\%$. Therefore, the low efficiencies are attributed to poor charge transport, resulting in a low η_{CC} (Fig. 1c). If charge collection is assisted by the application of an external voltage, a high η_{EQE} can be obtained¹⁷.

Growth of mixed layers at elevated substrate temperatures leads to phase separation and the appearance of crystalline domains. However, this increase in crystallinity and possibly larger L_{D} comes at the cost of an increased film roughness^{13,18}. The high density of pinholes leads to short circuits and makes device fabrication impractical. The same problem occurs when mixed-layer films are annealed post-deposition to induce phase separation.

Here, we present a method for the fabrication of bulk heterojunctions in small-molecule systems based on annealing mixed-layer films in a confined geometry. In this case, the devices are completed with a $\sim 1,000\text{-Å}$ -thick metal cathode, and then subsequently annealed. The metal cathode stresses the organic film during annealing, preventing morphological relaxation and the concomitant formation of a high density of pinholes, while

permitting phase separation to occur in the bulk of the organic film leading to the desired highly folded bulk heterojunction.

Annealing of completed devices was previously used to improve the lifetime¹⁹ and luminous efficiency^{20,21} of organic light-emitting diodes, and to improve the efficiency of polymer photovoltaic cells⁷.

However, these effects were not attributed to changes in the interface morphology but to polymer chain alignment, an increase in the charge carrier mobility or the formation of an interfacial layer at the organic/cathode interface.

Scanning electron microscope (SEM) images of the film surfaces

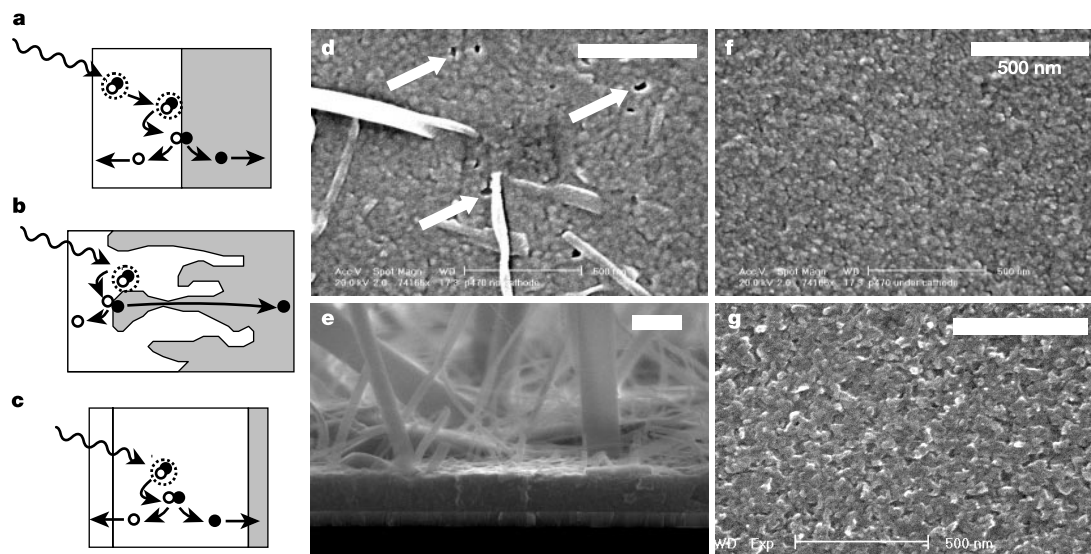


Figure 1 Diagrams of types of organic donor-acceptor photovoltaic cells and SEM images of the surface of a $\sim 5,000\text{-}\text{\AA}$ -thick CuPc:PTCBI film on ITO. **a–c**, Diagrams. **a**, Bilayer cell. **b**, Bulk heterojunction cell. **c**, Mixed-layer cell. In these diagrams, the incident light (wavy arrows), electrons (filled circles), holes (open circles) and excitons (paired circles) are indicated. **d–g**, SEM images. In **d** the film was annealed in the absence of a metal cap. White arrows indicate several pinholes. **e**, Cross-section of the same film obtained by

cleaving the substrate. **f**, The film was capped by a $1,000\text{-}\text{\AA}$ -thick film of Ag which was removed before imaging. For comparison, in **g** the organic surface of a non-annealed ITO/ $400\text{\AA}$ CuPc/ $400\text{\AA}$ PTCBI/ $1,000\text{\AA}$ Ag is shown after removal of the Ag cap. The features in this image correspond to crystalline domains of pure, planar-stacking PTCBI¹⁸. Scale bars, all 500 nm .

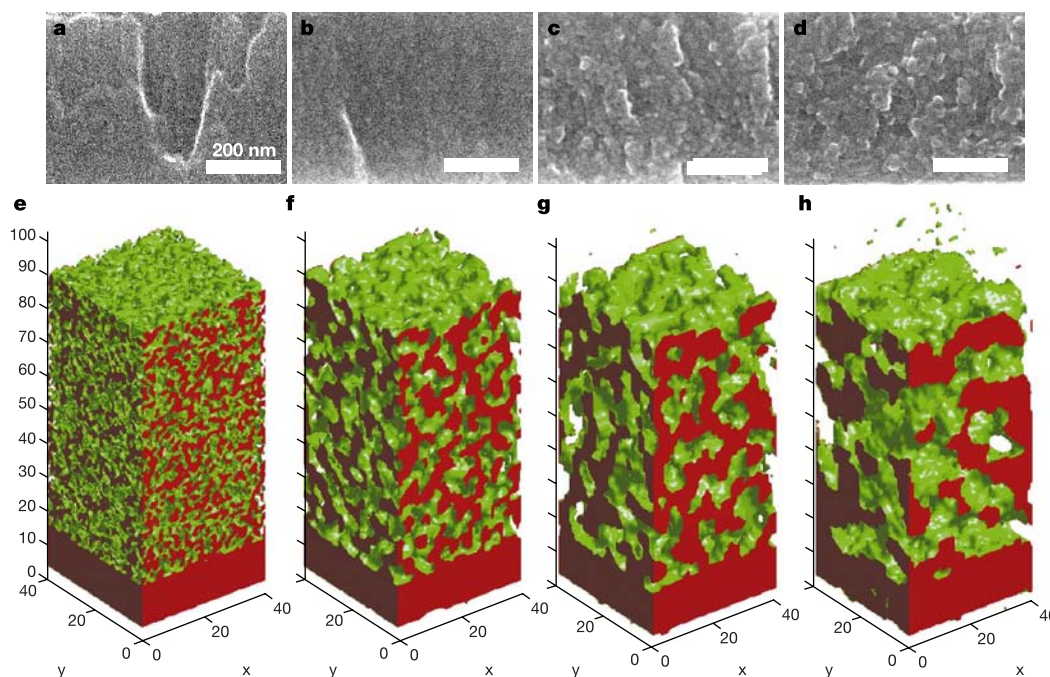


Figure 2 SEM images of cross-sections of a $5,000\text{-}\text{\AA}$ -thick CuPc:PTCBI(4:1) film on ITO. **a**, Not annealed. **b–d**, Annealed for **b**, 15 min at 450 K , **c**, 500 K and **d**, 550 K . **e–h**, The simulated effects of annealing on the interface morphology of a mixed-layer photovoltaic cell. The initial configuration (**e**) is generated using a random number generator, and assumes a mixture composition of 1:1. This also assumes that no significant phase segregation occurs during deposition. The interface between CuPc and PTCBI is shown as

a green surface. CuPc is shown in red and PTCBI is left 'transparent'. The as-grown, or initial, configuration is shown in **a**. The configurations after annealing at **f**, $T_{A1} = 0.067E_{\text{coh}}/N_{\text{AK}}$, **g**, $T_{A1} = 0.13E_{\text{coh}}/N_{\text{AK}}$ and **h**, $T_{A1} = 0.20E_{\text{coh}}/N_{\text{AK}}$ are also shown. Note the resemblance between the structure in images **a–d** and the simulated structures **e–h**. The vertical axis represents film thickness, and hence is the direction of film growth.

in Fig. 1d, e show the effect of capping by a 1,000-Å-thick Ag film during annealing. The layer structure was indium tin oxide (ITO)/100 Å copper phthalocyanine (CuPc)/600 Å CuPc:PTCBI (3:4)/100 Å 3,4,9,10-perylene tetracarboxylic bis-benzimidazole (PTCBI). The concentration of CuPc to PTCBI in this case was 3:4, by weight, achieved through codeposition. The images show the organic surface morphology after annealing for 2 min at 560 K. In Fig. 1d, e, the film was uncapped by metal during the annealing process, resulting in a high density of pinholes ($\sim 8 \times 10^8 \text{ cm}^{-2}$) and of large crystallites protruding from the film surface. In Fig. 1f, the organic layers were covered with a 1,000-Å-thick Ag cap during annealing that was subsequently peeled off using sticky tape. The resulting organic film is pinhole-free and lacks large ($\sim 1 \mu\text{m}$) crystalline domains, suggesting that the metal layer prevents morphological changes from occurring in the underlying film. For comparison, the surface morphology of a conventional non-annealed bilayer structure: ITO/400 Å CuPc/400 Å PTCBI/1,000 Å Ag after removing the Ag cap is shown in Fig. 1g.

While the presence of the Ag cap prevents the development of surface roughness, it does not prevent phase segregation within the bulk of the mixed layer itself. Evidence for phase segregation, leading to domains rich in CuPc or PTCBI, is provided in Fig. 2a–d. Here, SEM images of cross-sections of the layer structure: ITO/5,000 Å CuPc:PTCBI (4:1)/1,000 Å Ag are shown for an as-grown film (Fig. 2a), and for films annealed for 15 min at $T_{A1} = 450 \text{ K}$ (Fig. 2b), $T_{A1} = 500 \text{ K}$ (Fig. 2c), and $T_{A1} = 550 \text{ K}$ (Fig. 2d). The cross-section of the as-grown film (Fig. 2a) does not exhibit any morphological features other than artefacts of the cleaving process. Upon annealing, the cross-sections reveal domains whose size increases with increasing annealing temperature. At 550 K, domain sizes of $\sim 20 \text{ nm}$ are observed. This is corroborated by the X-ray diffraction data shown in Fig. 3. Upon annealing,

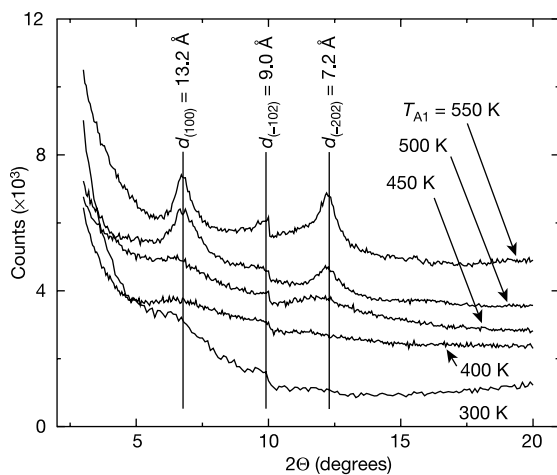


Figure 3 Bragg-Brentano X-ray diffractograms of a 5,000-Å-thick film on ITO using the Cu-K α line. The film was covered with a 1,000-Å-thick cap of Ag and annealed at 300 K (not annealed), and $T_{A1} = 400 \text{ K}$, 450 K, 500 K and 550 K. The Ag cap was removed before performing the scan. CuPc crystal indices are noted. The amorphous background is indicated by the broad curvature at low X-ray angles. The large width of the peaks suggests limited crystalline domain size. For the film annealed at 550 K, using the full-width at half-maximum of the peaks at angles $2\theta = 6.7^\circ$ and $2\theta = 12.2^\circ$, corresponding to interplanar spacings of $d = 13.2 \text{ Å}$ and $d = 7.2 \text{ Å}$, we calculate a domain size of $(12 \pm 1) \text{ nm}$, which is consistent with our observations in Fig. 2. Nevertheless, this represents a lower limit to the domain size, as the diffraction peaks are also broadened by molecular disorder and large strains associated with the growth of domains within an amorphous matrix. Another possible contribution to the peak width is the residual 'doping' of the CuPc and PTCBI-rich phases with PTCBI and CuPc, respectively.

diffraction peaks corresponding to the β -CuPc phase emerge²², and the broad amorphous background signal between $2\theta = 2.5^\circ$ and 12.5° is reduced.

In Fig. 2e–h, we modelled the effect of the annealing temperature T_{A1} on the interface morphology of a mixed-layer device using the cellular automaton model discussed in the Methods. Annealing at $T_{A1} = 0.067E_{\text{coh}}/N_A k$ (Fig. 2f), $T_{A1} = 0.13E_{\text{coh}}/N_A k$ (Fig. 2g) and $T_{A1} = 0.20E_{\text{coh}}/N_A k$ (Fig. 2h) has a dramatic influence on the morphology of the mixed-layer device, and the calculated morphology bears a remarkable resemblance to the observed cross-sections in Fig. 2a–d. Here E_{coh} is the cohesive energy of the crystalline organic solid per mole, N_A is Avogadro's number and k is Boltzmann's constant. Phase separation leads to the appearance of branches of pure material that grow increasingly thicker with increasing T_{A1} . The η_{ED} value is reduced in the thicker branches, but their presence improves η_{CC} .

By measuring the room-temperature η_{EQE} as a function of T_{A1} , the effect of this morphological change on exciton and charge transport can be inferred. In the inset of Fig. 4, the action spectrum of a mixed-layer device is shown as a function of T_{A1} . We observe a remarkable 30-fold increase in η_{EQE} at a wavelength of $\lambda = 690 \text{ nm}$ from 0.6% to 19%. The increase is uniform over the entire absorption spectrum of both CuPc and PTCBI, and cannot be identified with a single component. This confirms that the increase in η_{EQE} is associated with a change in morphology of the entire mixed layer.

In Fig. 4, η_{EQE} at $\lambda = 632 \text{ nm}$ as a function of T_{A1} is shown for a bilayer device (closed squares), and for mixed-layer devices (open symbols). Annealing a bilayer device does not significantly improve η_{EQE} , and annealing at $T_{A1} > 450 \text{ K}$ even results in its decrease. In contrast, for all mixed-layer devices, a dramatic increase in η_{EQE} is observed upon annealing at $T_{A1} > 450 \text{ K}$, with an optimal annealing temperature of $T_{A1} = 540 \text{ K}$. While the maximum attainable η_{EQE} clearly depends on the composition of the mixed layer, the η_{EQE} versus T_{A1} characteristics have a similar shape, independent of the mixture composition.

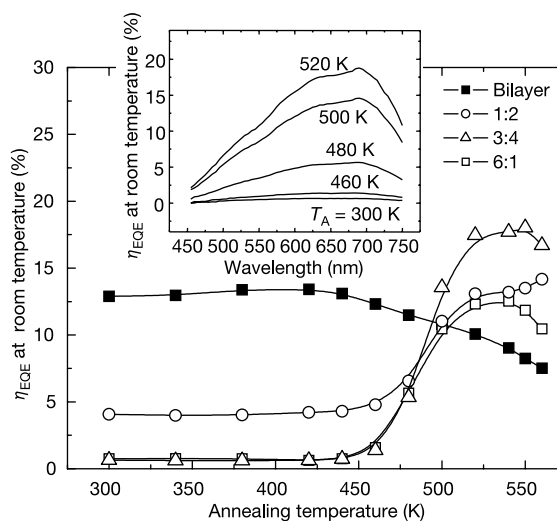


Figure 4 The room-temperature external quantum efficiency after annealing at various temperatures of bilayer and multi-layer devices. Closed squares, bilayer device with layer structure: ITO/400 Å CuPc/400 Å PTCBI/1,000 Å Ag. Open symbols, mixed-layer devices with layer structures: ITO/100 Å CuPc/600 Å CuPc:PTCBI ($x:y$)/100 Å PTCBI/1,000 Å Ag, where $x:y$ is 1:2 (circles), 3:4 (triangles) and 6:1 (open squares). The cells were subsequently annealed for 2 min at 340 K and 380 K, then every 20 K between 420 K and 540 K, and 550 K and 560 K, each time returning to room temperature between annealing steps to measure η_{EQE} . Inset, Room-temperature η_{EQE} after annealing at various temperatures of a device with layer structure: ITO/100 Å CuPc/600 Å CuPc:PTCBI (3:4)/100 Å PTCBI/1,000 Å Ag. The cell was annealed and measured as in the main figure.

Table 1 Effect of treatments on the properties of a mixed-layer solar cell

	J_{SC} ($\mu A cm^{-2}$)	V_{OC} (V)	FF	η_P (%)
As-grown bilayer	340	0.33	0.52	0.75 ± 0.05
As-grown mixed layer	15.5	0.26	0.25	$(1.3 \pm 0.1) \times 10^{-2}$
First anneal ($T_{A1} = 520$ K)	190*	0.10	0.27	$(6.5 \pm 0.4) \times 10^{-1}$
Contact replacement	250	0.30	0.26	0.25 ± 0.2
Second anneal ($T_{A2} = 500$ K)	880	0.44	0.31	1.5 ± 0.1

Shown are the room-temperature performance characteristics of a ITO/400 Å CuPc/400 Å PTCBI/1,000 Å Ag bilayer and ITO/100 Å CuPc/600 Å CuPc:PTCBI (6:1)/100 Å PTCBI/1,000 Å Ag mixed-layer solar cell. The illumination source is a tungsten-halogen lamp with a power density of $7.8 mW cm^{-2}$. Here, J_{SC} is the short-circuit current density.

*This result for J_{SC} is in contrast to the results for η_{EQE} of a device with an identical layer structure (Fig. 4), where η_{EQE} of the annealed mixed-layer device approaches that of the as-grown bilayer device. This apparent contradiction is a consequence of the higher optical power levels used during measurements of the current-voltage characteristics as compared to the η_{EQE} measurements.

The temperature-dependent η_{EQE} data also allow us to estimate the maximum operating lifetime (t_{op}) of the devices employing these structurally metastable materials. For example, for $t_{op} = 10$ yr at a cell temperature T_{op} , we obtain $T_{op} = (k \ln(t_{op}/t_{A1})/\Delta E_A + T_{A1}^{-1})^{-1} \approx 53^\circ C$ where t_{A1} is the annealing time (~ 1 min) at $T_{A1} = 540$ K required to effect the observed morphological changes, and $\Delta E_A = (1.1 \pm 0.1)$ eV is the activation energy characteristic of the annealing process estimated from the data in Fig. 4. Phase separation during normal operation of the cells is therefore not expected to contribute significantly to degradation of cell performance.

In Table 1 the room-temperature characteristics of a mixed-layer device are listed as a function of the annealing treatment. For

reference, the performance parameters of a bilayer device are also shown. Prior to annealing, the short-circuit current density ($J_{SC}^M = 15.5 \mu A cm^{-2}$) of the mixed-layer cell is more than an order of magnitude smaller than that of the bilayer ($J_{SC}^B = 340 \mu A cm^{-2}$), leading to a low $\eta_P = (1.3 \pm 0.1) \times 10^{-2}\%$. After annealing at $T_{A1} = 520$ K, $J_{SC}^M = 190 \mu A cm^{-2}$. The drop in V_{OC} from 0.26 V to 0.10 V partially offsets the gains in J_{SC} , leading to $\eta_P = (6.5 \pm 0.4) \times 10^{-1}\%$.

The drop in V_{OC} is attributed to an increased resistance arising from a reduction in disorder²³ at the organic/Ag interface that is due to the annealing process. Hence, improvements in performance are achieved by replacing the contact by peeling off the 'confining' Ag layer and replacing it by deposition of a 120 Å BCP/1,000 Å Ag contact. This contact replacement results in an increased $J_{SC}^M = 250 \mu A cm^{-2}$ and $V_{OC} = 0.30$ V (see Table 1). Annealing this device a second time at $T_{A2} = 500$ K further improves J_{SC}^M to $880 \mu A cm^{-2}$. The open circuit voltage of $V_{OC}^M = 0.44$ V also exceeds that of the bilayer device ($V_{OC}^B = 0.33$ V). The η_P of the twice-annealed mixed-layer device with a replaced contact is $\eta_P = (1.5 \pm 0.1)\%$, a twofold improvement over the bilayer with $\eta_P = (0.75 \pm 0.05)\%$.

The contact replacement strategy was used to fabricate a solar cell with a high η_P value under standard AM1 illumination conditions at an intensity of $105 mW cm^{-2}$ (that is, ~ 1 Sun). The device layer structure: ITO/150 Å CuPc/440 Å CuPc:PTCBI (1:1)/100 Å PTCBI/1,000 Å Ag was first annealed at $T_{A1} = 520$ K for 2 min. The contact was subsequently peeled off and replaced by deposition of a 150 Å BCP/1,000 Å Ag contact. The solar cell performance characteristics after the second anneal are shown in Fig. 5a as a

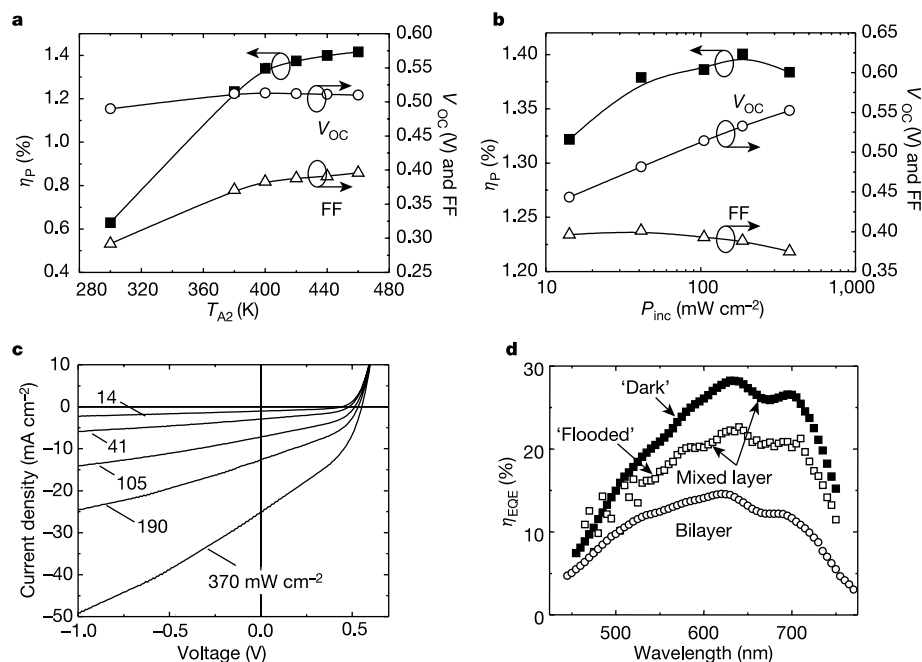


Figure 5 Parameters affecting the room-temperature power conversion efficiency.

a, η_P , the open-circuit voltage, V_{OC} , and the fill factor, FF, as functions of the second annealing temperature, T_{A2} , for the layer structure: ITO/150 Å CuPc/440 Å CuPc:PTCBI (1:1)/100 Å PTCBI/150 Å BCP/1,000 Å Ag, where the BCP/Ag layers were deposited after the first anneal (at $T_{A1} = 520$ K). The second annealing process is essentially complete at $T_{A2} = 400$ K, so the mechanism leading to cell improvement is different from that of the first annealing step. We infer that the role of the second annealing process is to remove contaminants such as H_2O or O_2 from the donor-acceptor interfaces, which provide sites for exciton and/or charge recombination. Indeed, a similar increase in η_P is observed when a sample that was exposed to air after the first anneal was annealed a second time. Air exposure causes a rapid decrease in η_P , reducing it to less than 50% of the pre-exposure value. Here, the pre-exposure η_P is recovered after annealing to 400 K.

It is possible that some 'forming' of the donor-acceptor mixed layer/BCP contact also occurs during the second thermal treatment. **b**, Room-temperature η_P , V_{OC} and FF, as functions of the incident optical power intensity P_{inc} , after the second annealing process at $T_{A2} = 460$ K for the same layer structures as in **a**. The photocurrent has a linear dependence on the illumination intensity as shown in **c**, and the increase in V_{OC} with increased illumination intensity offsets the decrease in FF, resulting in η_P being nearly independent of the illumination intensity. **c**, Room-temperature current density-voltage characteristic of the device of **b** at various incident power levels. **d**, External quantum efficiency, η_{EQE} , of the mixed-layer device of **b**, measured with (open squares) and without (closed squares) flooding by $105 mW cm^{-2}$ AM1 illumination. For comparison, the η_{EQE} of an optimized ITO/200 Å CuPc/200 Å PTCBI/150 Å BCP/Ag bilayer structure is also shown (open circles).

function of T_{A2} . A maximum efficiency is reached for $T_{A2} = 460$ K, with $\eta_p = (1.42 \pm 0.07)\%$ representing the highest efficiency (by $\sim 50\%$) achieved for CuPc/PTCBI photovoltaic 'Tang' cells over the last 16 yr (ref. 1).

The dependence of the performance characteristics of this device on the incident optical power is shown in Fig. 5b. The slightly lower η_p at 105 mW cm^{-2} as compared to Fig. 5a is due to minor device degradation. In Fig. 5c, the current-voltage characteristics are also shown as a function of intensity. At -1 V bias, the photocurrent density is approximately twice that obtained under short-circuit conditions. This strong dependence of photocurrent on applied bias suggests that carrier collection ultimately limits η_p . Optimizing this η_{CC} may lead to $>$ twofold improvements in J_{SC} , and hence η_p .

In Fig. 5d, η_{EQE} is shown, measured with (open squares) and without (filled squares) flooding by 105 mW cm^{-2} AM1 white-light illumination. For comparison, η_{EQE} of an optimized bilayer device: ITO/200 Å CuPc/200 Å PTCBI/150 Å BCP/Ag (ref. 24) is also shown (open circles). The peak 'dark' $\eta_{EQE}^M = 28\%$ of the annealed mixed-layer device is twice that of the bilayer device $\eta_{EQE}^B = 14\%$. The decrease in η_{EQE} upon flooding with white light is a consequence of the increased carrier concentration under illumination, which increases the recombination probability, and hinders charge transport because of space-charge build-up within the complex folds of the bulk heterojunction structure.

The observed efficiency enhancement relies on the annealing of mixed-layer films in a confined geometry, that is, with a metal contact that prevents stress relief during morphological relaxation that typically occurs in molecular materials at elevated temperatures. Measurements on mixed-layer devices after annealing show dramatic increases in their external quantum efficiencies by a factor of up to ~ 30 . We anticipate that further understanding of the nanoscale structure of confined materials, and the aspects of phase separation of binary mixtures of small-molecular-weight materials observed here, will facilitate the realization of high-performance organic photovoltaic and other optoelectronic devices. $\square$

Methods

The photovoltaic cells were deposited on glass substrates pre-coated with a $1,500\text{-\AA}$ -thick, transparent, conducting ITO anode (sheet resistance 40Ω per square). The substrates were cleaned immediately before transferring them into the vacuum system for film deposition²⁵. The organic materials were commercially obtained and purified before deposition using thermal gradient sublimation¹⁸. The photoactive materials used were CuPc and PTCBI, and bathocuproine (BCP) was used as a contact layer. The organic layers were grown by high vacuum thermal evaporation (base pressure 10^{-7} – 10^{-6} torr) from a tungsten 'boat' onto a room-temperature substrate. This was followed by the deposition of the metal cathode through a shadow mask, resulting in contact diameters of 0.3 mm and 1 mm. No dependence of the device characteristics on device area was observed.

After fabrication, the cells were transferred to a vacuum chamber held at 30 mtorr with a heating stage, electrical probes and windows for optical access. The temperature ramp rate of the heating stage was fixed at $15^\circ \text{C min}^{-1}$. Electrical characterization was performed during annealing using a semiconductor parameter analyser to obtain the current-voltage characteristics. For *in situ* photovoltaic power efficiency measurements, the devices were illuminated through the substrate with a $1,000\text{-W}$ solar simulator (Oriol Corp.) equipped with an AM1 filter. To measure η_{EQE} , a monochromatic beam of variable wavelength light chopped at 400 Hz (50% duty cycle) was focused onto a 1-mm diameter device. The photocurrent was measured using a lock-in amplifier referenced to the chopper frequency.

To gain a better understanding of the underlying physical process of the phase separation on the performance of mixed-layer photovoltaic cells, we implemented a model using cellular automata. This approach provides a numerically efficient and at the same time phenomenologically sound method of discretely simulating recrystallization and grain growth²⁶. Briefly, a volume is discretized into a three-dimensional array in a simple cubic lattice containing $N_x \times N_y \times N_z = N$ cells. We define the z -direction as the growth direction (that is, perpendicular to the substrate plane). Periodic boundary conditions are applied in the x and y directions. The free energy of a configuration is:

$$E = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^6 E_{M(i), M(j)} \quad (1)$$

where j sums over all nearest neighbours, $M(i)$ is the material at location i , and $E_{A,B}$ is the free energy associated with the molecular contact between molecules A and B . In this scheme, the cohesive energy per mole of material A is $E_{coh} = 3N_A E_{A,A}$, where N_A is Avogadro's number. E_{coh} is the evaporation enthalpy which can be obtained by thermogravimetry. In our simulations, only two materials CuPc

($E_{coh, \text{CuPc}} = 176 \text{ kJ mole}^{-1}$; ref. 27) and PTCBI are used. Since $E_{coh, \text{PTCBI}}$ is unknown, and since most small molecular organic materials used in organic electronic devices have similar E_{coh} values²⁷, we assume for convenience that $E_{PTCBI} = E_{CuPc}$. Furthermore, we assume that $2E_{CuPc, \text{PTCBI}} = E_{CuPc, \text{CuPc}} = E_{PTCBI, \text{PTCBI}}$ (ref. 28).

The lattice is initialized to mimic the as-grown mixed structures. Subsequently, phase segregation is modelled using a single transformation rule: two neighbouring molecules can exchange positions. Assuming that R_0 is the rate at which molecular exchanges are attempted per cell, the rate of attempts able to overcome the energy barrier, ΔE_A , of exchanging two molecules is a function of temperature: $R(T) = R_0 \exp(-\Delta E_A/kT)$, where k is Boltzmann's constant and T is the absolute temperature. The activation energy associated with the switching of two molecules is prohibitively high because it would require the molecules to deform significantly. The actual process thus involves the presence of vacancies whose activation energy is that responsible for the generation of those vacancies.

Received 25 April; accepted 23 July 2003; doi:10.1038/nature01949.

1. Tang, C. W. A two-layer organic solar cell. *Appl. Phys. Lett.* **48**, 183–185 (1986).
2. Yu, G., Zhang, C. & Heeger, A. J. Dual-function semiconducting polymer devices: light-emitting and photodetecting diodes. *Appl. Phys. Lett.* **64**, 1540–1542 (1994).
3. Halls, J. J. M. et al. Efficient photodiodes from interpenetrating networks. *Nature* **376**, 498–500 (1995).
4. Granström, M. et al. Laminated fabrication of polymeric photovoltaic diodes. *Nature* **395**, 257 (1998).
5. Yu, G., Gao, J., Hummelen, J., Wudl, F. & Heeger, A. J. Polymer photovoltaic cells—enhanced efficiencies via a network of internal donor-acceptor heterojunctions. *Science* **270**, 1789–1791 (1995).
6. Peumans, P., Yakimov, A. & Forrest, S. R. Small molecular weight organic thin-film photodetectors and solar cells. *J. Appl. Phys.* **93**, 3693–3723 (2003).
7. Padinger, E., Rittberger, R. S. & Sariciftci, N. Effects of postproduction treatment on plastic solar cells. *Adv. Funct. Mater.* **13**, 85–88 (2003).
8. Shaheen, S. E. et al. 2.5% efficient organic plastic solar cells. *Appl. Phys. Lett.* **78**, 841–843 (2001).
9. Aernouts, T. et al. Extraction of bulk and contact components of the series resistance in organic bulk donor-acceptor heterojunctions. *Thin Solid Films* **403–404**, 297–303 (2002).
10. Feng, W., Fujii, A., Lee, S., Wu, H. & Yoshino, K. Broad spectral sensitization of organic photovoltaic heterojunction device by perylene and C_{60} . *J. Appl. Phys.* **88**, 7120–7123 (2000).
11. Dittmer, J. J. et al. Crystal network formation in organic solar cells. *Sol. Energy Mater. Sol. Cells* **61**, 53–61 (2000).
12. Tsuzuki, T., Shirota, Y., Rostalski, J. & Meissner, D. The effect of fullerene doping on photoelectric conversion using titanyl phthalocyanine and a perylene pigment. *Sol. Energy Mater. Sol. Cells* **61**, 1–8 (2000).
13. Geens, W., Aernouts, T., Poortmans, J. & Hadzioannou, G. Organic co-evaporated films of a PPV-pentamer and C-60: model systems for donor/acceptor polymer blends. *Thin Solid Films* **403–404**, 438–443 (2002).
14. Hiramoto, M., Suemori, K. & Yokoyama, M. Photovoltaic properties of ultramicrostructure-controlled organic co-deposited films. *Jpn. J. Appl. Phys.* **41**, 2763–2766 (2002).
15. Halls, J. J. M. & Friend, R. H. The photovoltaic effect in a poly(p-phenylenevinylene)/perylene heterojunction. *Synth. Metals* **85**, 1307–1308 (1997).
16. Peumans, P. & Forrest, S. R. Very-high-efficiency double-heterostructure copper phthalocyanine/ C_{60} photovoltaic cells. *Appl. Phys. Lett.* **79**, 126–128 (2001).
17. Peumans, P., Bulovic, V. & Forrest, S. R. Efficient, high-bandwidth organic multilayer photodetectors. *Appl. Phys. Lett.* **76**, 3855–3857 (2000).
18. Forrest, S. R. Ultrathin organic films grown by organic molecular beam deposition and related techniques. *Chem. Rev.* **97**, 1793–1896 (1997).
19. Kim, J., Lee, J., Han, C., Lee, N. & Chung, I.-J. Effect of thermal annealing on the lifetime of polymer light-emitting diodes. *Appl. Phys. Lett.* **82**, 4238–4240 (2003).
20. Niu, Y., Hou, Q., Yuan, M. & Cao, Y. Effects of thermal annealing on light-emitting devices based on fluorene-copolymers with thiophene and ethylenedioxythienylene. *Synth. Met.* **135–136**, 477–478 (2003).
21. Lee, T.-W. & Park, O. O. Effect of electrical annealing on the luminous efficiency of thermally annealed polymer light-emitting diodes. *Appl. Phys. Lett.* **77**, 3334–3336 (2000).
22. Prabakaran, R., Kesavamoorthy, R., Reddy, G. & Xavier, F. Structural investigation of copper phthalocyanine thin films using X-ray diffraction, Raman scattering and optical absorption measurements. *Phys. Stat. Sol.* **229**, 1175–1186 (2002).
23. Baldo, M. A. & Forrest, S. R. Interface-limited injection in amorphous organic semiconductors. *Phys. Rev. B* **64**, 085201 (2001).
24. Peumans, P., Bulovic, V. & Forrest, S. R. Efficient photon harvesting at high optical intensities in ultrathin double-heterostructure photovoltaic diodes. *Appl. Phys. Lett.* **76**, 2650–2652 (2000).
25. Burrows, P. E. et al. Relationship between electroluminescence and current transport in organic heterojunction light-emitting devices. *J. Appl. Phys.* **79**, 7991–8006 (1996).
26. Raabe, D. "Cellular automata in materials science with particular reference to recrystallization simulation". *Annu. Rev. Mater. Res.* **32**, 36–76 (2002).
27. Shtein, M., Gossenberger, H. F., Benziger, J. B. & Forrest, S. R. Material transport regimes and mechanisms for growth of molecular organic thin films using low-pressure organic vapor phase deposition. *J. Appl. Phys.* **89**, 1470–1476 (2001).
28. Israelachvili, J. N. *Intermolecular and Surface Forces* 176–212 (Academic, London, 1992).

Acknowledgements We thank N. Stroustrup for his assistance, and gratefully acknowledge the financial support of the Air Force Office of Scientific Research, the National Renewable Energy Laboratory, and the Global Photonic Energy Corporation. S.U. is currently on leave from Nippon Oil Corporation, Central Technical Research Laboratory, 8 Chidori-cho, Naka-ku, Yokohama 231-0815, Japan.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.R.F. (forrest@princeton.edu).

Lubrication by charged polymers

Uri Raviv¹, Suzanne Giasson², Nir Kampf¹, Jean-François Gohy³, Robert Jérôme³ & Jacob Klein^{1,4}

¹Weizmann Institute of Science, Rehovot 76100, Israel

²Department of Chemistry and School of Pharmacy, University of Montreal, Québec, Canada H3C 3J7, and CERSIM, Laval University, Québec, Canada G1K 7P4

³Center for Education and Research on Macromolecules, University of Liège, Sart-Tilman B6, 4000 Liège, Belgium

⁴Physical and Theoretical Chemistry Laboratory, Oxford University, South Parks Road, Oxford OX1 3QZ, UK

Long-ranged forces between surfaces in a liquid control effects from colloid stability¹ to biolubrication², and can be modified either by steric factors due to flexible polymers³, or by surface charge effects⁴. In particular, neutral polymer 'brushes' may lead to a massive reduction in sliding friction between the surfaces to which they are attached^{5–7}, whereas hydrated ions can act as extremely efficient lubricants between sliding charged surfaces⁸. Here we show that brushes of charged polymers (polyelectrolytes) attached to surfaces rubbing across an aqueous medium result in superior lubrication compared to other polymeric surfactants. Effective friction coefficients with polyelectrolyte brushes in water are lower than about 0.0006–0.001 even at low sliding velocities and at pressures of up to several atmospheres (typical of those in living systems). We attribute this to the exceptional resistance to mutual interpenetration displayed by the compressed, counterion-swollen brushes, together with the fluidity of the hydration layers surrounding the charged, rubbing polymer segments. Our findings may have implications for biolubrication effects, which are important in the design of lubricated surfaces in artificial implants, and in understanding frictional processes in biological systems.

Normal and shear forces between mica surfaces immersed in water were measured using a surface force balance (SFB) described earlier⁹ (Fig. 1 bottom inset). Control measurements were carried out in salt-free water, revealing the long-ranged repulsion due to counterion osmotic pressure followed by a jump, on approach, into adhesive contact, in agreement with earlier observations^{10,11}. The mica surfaces were then made hydrophobic ('hydrophobized'), as described elsewhere¹², by coating with a layer of stearic trimethylammonium iodide (STAI: CH₃(CH₂)₁₇N⁺(CH₃)₃I[–]), and on approach across water jumped into a strongly adhesive contact due to the hydrophobic attraction between the STAI layers¹³ (Fig. 1 top inset). Finally, polyelectrolyte (PE) brushes were created on the hydrophobized mica using the diblock copolymer poly(methyl methacrylate)-*block*-poly(sodium sulphonated glycidyl methacrylate) copolymer (PMMA-*b*-PSGMA), anionically synthesized and characterized as reported in detail elsewhere¹⁴. The copolymer is a hydrophobic (PMMA)–hydrophilic (polyelectrolytic PSGMA) diblock: (CH₂–C(CH₃)CO₂CH₃)₄₁(CH₂–C(CH₃)CO₂–CH₂CHOHCH₂SO₃[–]Na⁺)₁₁₅, where the polyelectrolytic block is 70% sulphonated. Control force versus distance measurements showed that, as expected, this negatively charged polymer does not adhere from aqueous solution to the bare negatively charged mica surface. Following hydrophobization of the surfaces and their incubation with the diblock, subsequent force profiles (Fig. 1) reveal clearly the attachment of the chains to the mica surfaces. The attachment must be via their hydrophobic PMMA moieties to form end-tethered layers of charged PSGMA moieties, as under these conditions the polyelectrolyte block, PSGMA, does not itself adsorb on hydrophobic surfaces¹⁵. The unperturbed thickness *L* of the PE brush is estimated from the profiles in Fig. 1 as *L* = 13 ± 2 nm. Refractive index measurements on the confined

PE layers revealed an adsorbance $\Gamma = 3 \pm 1 \text{ mg m}^{-2}$ of the polymer on the hydrophobized mica, corresponding to a mean surface spacing between chains of $4 \pm 0.7 \text{ nm}$.

The PE-brush-bearing mica surfaces were then progressively compressed, and the shear or frictional forces *F_s* between them measured as the top surface was made to slide back and forth past the lower one at increasing loads *F_n* and at different sliding velocities *v_s*. Typical results are shown in Fig. 2. Trace a in Fig. 2 shows the periodic lateral motion applied to the top mica surface, while traces b–d, taken directly from the recording oscilloscope, show the shear force transmitted to the lower surface across the interacting PE brushes at increasing compressions. The right inset in Fig. 2 shows the frequency dependence of the *F_s* responses in traces b–d, with arrows marking the signal at the drive frequency. Within the resolution of our apparatus, no systematic forces at the drive frequency (0.5 Hz for the traces shown) were measurable above the scatter $\delta F_s = \pm 20\text{--}30 \text{ nN}$, up to loads corresponding to a mean pressure of a few atmospheres. We may define an effective friction coefficient $\mu_{\text{eff}} = F_s/F_n$, where *F_s* is the force required to slide the surfaces past each other under a normal load *F_n*. At the higher loads before chain removal, where the mean pressure between the surfaces is $\sim 3 \times 10^5 \text{ Pa}$, (as in trace c of Fig. 2, typical of other traces at these loads), *F_s* ≤ δF_s and we find $\mu_{\text{eff}} \lesssim 0.0006$. Within the range of

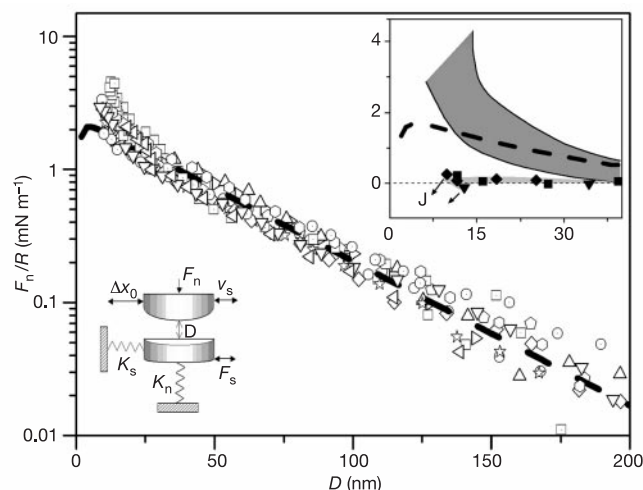


Figure 1 Force versus separation profiles. Normalized force ($F_n(D)/R$) versus surface separation (*D*) profiles between hydrophobized (STAI-coated) mica surfaces bearing PMMA-*b*-PSGMA diblock copolymers tethered by their PMMA moiety, where *R* is the mean radius of curvature of the surfaces. *D* = 0 refers to contact between bare mica surfaces in purified water. Profiles are taken following overnight incubation in $100 \pm 10 \mu\text{g ml}^{-1}$ PMMA-*b*-PSGMA aqueous solution. Results from different sets of experiments (different pairs of mica sheets) and different contact positions in each experiment, measured on first and second compression–decompression cycles, are shown. The broken curve is a fit to the DLVO expression²⁰ $F_n(D)/R = 128\pi c k_B T \kappa^{-1} \tanh^2(e\psi_0/4k_B T) \exp(-\kappa D) - A_H/6D^2$, where *c* is the ion concentration, *T* is the temperature (296 K), *k_B* is Boltzmann's constant, *A_H* is the Hamaker constant ($= 1.5 \times 10^{-20} \text{ J}$, the theoretical value for the interaction of two hydrocarbon-coated mica surfaces across water⁴), ψ_0 is the effective (far-field) surface potential, κ^{-1} is the Debye length ($40 \pm 3 \text{ nm}$, corresponding to a 1:1 ion concentration $c = (6 \pm 1) \times 10^{-5} \text{ M}$), and *e* is the electronic charge. The upper inset shows the data from the main figure at short separations on a larger scale (as a shaded band); from the deviation of the broken line from the data we estimate an unperturbed brush thickness of $13 \pm 2 \text{ nm}$ on each surface. The data points in the inset are the forces between the STAI-coated mica surfaces across water before incubation in the PMMA-*b*-PSGMA solution, showing the jump (at *J*) into strong adhesive contact. The lower inset shows schematically the SFB configuration, where *K_s* and *K_n* are the constants of the shear- and normal-force springs, respectively.

shear velocities studied, $250 \text{ nm s}^{-1} < v_s < 500 \text{ nm s}^{-1}$, the shear forces were, in all cases, within the noise level δF_s down to $D \approx 10 \text{ nm}$. For this reason we could not determine a velocity or shear-rate dependence of the shear force, though the behaviour of the confined film in this regime was clearly liquid-like in the sense of not sustaining a yield stress. At the highest compressions, where the volume fraction of the compressed polymer was close to unity (for example, trace d in Fig. 2), a small shear force could be detected above the scatter just before the jump at J, from $D = 5.8 \pm 0.3 \text{ nm}$ into strong adhesive contact at $D = 1.9 \pm 0.3 \text{ nm}$. In the adhesive contact following the jump the remaining confined film showed solid-like behaviour, in that a strong shear force was required before yield (sliding) was observed (right-hand side of trace d in Fig. 2).

We attribute this jump to an abrupt removal of the PE brush layers. As the top surface slides back and forth laterally, the surfaces approach slowly under thermal drift. The frictional drag between them grows, and some chains experience a force sufficient to overcome their hydrophobic adhesion to the STAI layers, causing them to detach. The shear force per chain on the remaining chains then becomes correspondingly larger, causing more of them to detach, and a rapid catastrophic removal of the diblocks occurs, leading to the jump observed at the point J (Fig. 2). It is of interest that the separation to which the surfaces jump is comparable with the double thickness of the hydrophobizing STAI layers^{12,13}, indicating the shear-off from the intersurface gap of most of the confined diblock. Such a shearing away of the PE brushes could be overcome by a more tenacious end-attachment of the PE moieties, such as by covalent tethering of the brush-chains to each surface, or endocytic anchoring for the case of PE layers at the outer surface of cell membranes. If the surfaces are separated

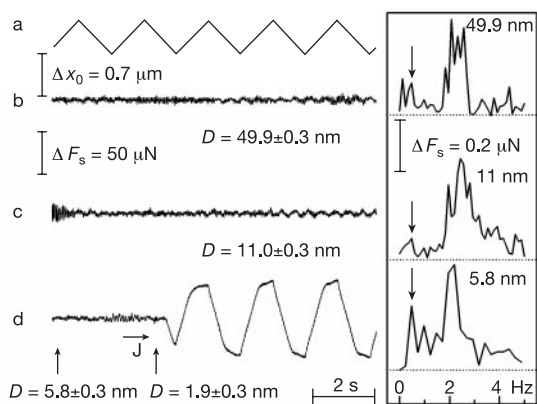


Figure 2 Frictional forces between PE brushes. Main panel; traces show the shear forces F_s between STAI-coated mica surfaces bearing the PMMA-*b*-PSGMA brushes in conductivity water, as they slide past each other at different surface separations D , and are taken directly from the recording oscilloscope. Trace a shows the lateral back-and-forth lateral motion (Δx_0 , lower inset to Fig. 1) applied to the top mica surface. Traces b–d show the corresponding shear force F_s transmitted between the surfaces at the surface separations indicated. Frequency analysis (inset, right) reveals that the magnitude of F_s at the frequency of the lateral drive motion (0.5 Hz, indicated by arrows) for a surface separation $D = 11 \text{ nm}$ (trace c) is, within the noise-limiting sensitivity ($\delta F_s = \pm 30 \text{ nN}$), indistinguishable from its value at large separations (trace b, $D = 49.9 \text{ nm}$). In trace d the surfaces have been further compressed and, just before jump-in at the point J, from $D = 5.8 \pm 0.3 \text{ nm}$ to $D = 1.9 \text{ nm}$ (when the PE brushes are sheared off, see text), the shear force at the drive frequency increases to $F_s(0.5 \text{ Hz}) = 200 \pm 30 \text{ nN}$. The mean pressure between the surfaces is given by $P = F_n(D)/A$, where F_n is the load and the area A may be evaluated from hertzian contact mechanics⁴. For $D = 5.8 \text{ nm}$ (trace d before jump-in), $P = 3 \times 10^5 \text{ N m}^{-2}$. Very similar traces to d (not shown) were obtained in other runs just before shear-off of the brushes.

from the adhesive contact following the PE shear-off, both the normal force profile and the very low friction, as well as the adsorbance revealed by refractive index measurements, recover to their equilibrium values (as in Figs 1 and 2, respectively) within 30 min.

These low values of the effective friction coefficient μ_{eff} reveal a remarkable capacity for lubrication by the PE brushes. We attribute this to two main factors. On the one hand, mutually compressed polymer brushes in a good solvent resist interpenetration, owing to excluded volume effects arising from chain configurational entropy^{3,6,16–18}. Additionally, for the case of the PE brushes this configurational excluded volume effect is augmented by the large osmotic pressure exerted by mobile counterions within the brush^{19–23}, whose concentration, even in the unperturbed PE brushes, is $\sim 0.3 \text{ M}$. This implies that a given load can be supported with less mutual interpenetration than for neutral brushes²⁰, thereby reducing the extent of the sheared interfacial region when the opposing surfaces slide past each other. Indeed, an estimate of the extent of mutual interpenetration of the brushes in our system (U.R. *et al.*, manuscript in preparation), based on the arguments of ref. 16 together with the calculations in ref. 22, suggests this is no more than $\sim 1 \text{ nm}$ at the highest compressions before shear-off of the brushes. At the same time, each of the charged PE segments rubbing against others within the sheared interpenetration zone is surrounded by a hydration sheath. As has recently been demonstrated, such sheaths are tenaciously bound to the charges but are at the same time very fluid, and so serve as efficient lubricating layers⁸.

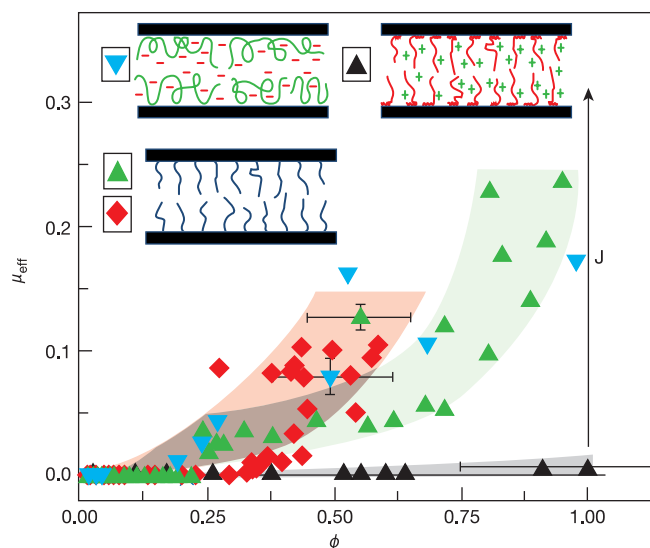


Figure 3 Variation of the effective friction coefficient μ_{eff} with volume fraction ϕ of confined polymer for different polymer lubricants. Volume fractions are based on absolute adsorbance values determined for the respective polymers from *in situ* refractive index measurements. Red and green symbols are for neutral brushes in non-polar⁵ and in aqueous²⁴ good solvents respectively, with respective bands indicating the range of the scatter. Blue symbols are for an adsorbed cationic polyelectrolyte, chitosan, in an aqueous solution at pH 3.5 (ref. 31); the normal force profiles of this chitosan sample, $M = 6 \times 10^5$ and degree of deacetylation 85% (Fluka), are very similar to those reported in ref. 28 where a comparable chitosan sample was used in similar conditions. The adsorbance of our chitosan sample onto each mica surface is 1.2 mg m^{-2} , determined by *in situ* refractive index measurements. The data shown are from two independent experiments and different contact points within each experiment. The black symbols and corresponding grey band are from the present study on the PMMA-*b*-PSGMA brushes. Shear velocities for all data are in the range $250\text{--}500 \text{ nm s}^{-1}$. The cartoons illustrate the different charged and uncharged polymer lubricant configurations, with positive or negative charge signs indicating the counterions.

Refinement of the above model will require more detailed investigation of how the shear and normal forces vary with PE length, charge density, adsorbance and attachment to the surface, as well as with salt concentration in the surrounding aqueous medium.

As this combination of charge and steric effects is absent in most other forms of polymeric surfactant, it is instructive to compare the frictional effects for different classes of solvated polymer surface layers. These include neutral brushes in both aqueous²⁴ and organic^{5–7} liquids, and adsorbed (as opposed to end-tethered) neutral²⁵ and charged polymers. Many different measures for such a comparison are clearly possible, even when limiting ourselves to frictional studies using the standard SFB configuration. A particularly useful measure at high compressions—the most interesting regime—which acts to normalize effects due to the different surface architectures, surface densities and layer thicknesses, is to examine how μ_{eff} varies with the volume fraction ϕ of the compressed polymer layers. This is because volume fractions of the compressed polymers are directly related to the pressure between the surfaces, but primarily because at the high volume fractions corresponding to the strongest compressions, these pressures become independent of the details of the polymer conformations²⁶. The osmotic pressures supporting the load between the rubbing surfaces then depend—in a good solvent—on the monomer concentrations²⁶. At the same time, for an unentangled interpenetration zone, the viscous dissipation within the sheared layer depends—for a given shear rate and concentration—on the monomeric friction alone^{16,27}.

In Fig. 3 we compare the present results on PE brushes with frictional SFB studies on other generic cases, at comparable shear velocities: neutral brushes in an organic solvent⁵ and in water²⁴, and an adsorbed charged polymer (chitosan) in aqueous solution²⁸ (N.K., U.R. and J.K., unpublished data), all in good solvent conditions. The effective friction coefficient μ_{eff} defined earlier is plotted against the volume fraction ϕ of the compressed, sheared polymers. All polymer surfactants lead to very low effective friction coefficients at volume fractions $\phi \lesssim 0.3$. But whereas the neutral brushes and the adsorbed charged chains (coloured symbols and bands in Fig. 3) show a rapid increase in μ_{eff} at higher volume fractions, friction mediated by PE brushes (black symbols and grey band) remains extremely low up to the point of brush removal J (by which point the volume fraction is, within the scatter, close to unity). For the neutral brushes in aqueous and organic solvents^{5–7,24}, this increase has been attributed largely to the sharp increase in the effective viscosity of the sheared interfacial region at the higher ϕ values, together with the greater mutual interpenetration of the opposing layers at the higher loads. For the adsorbed PE chain (blue symbols in Fig. 3), where counterion osmotic effects and lubricating hydration sheaths are expected to be similar to those acting in the PE brush, the origin of the rapid increase in the friction may be different, and associated rather with bridging effects. This effect, suggested also to explain the higher frictional drag between neutral adsorbed chains relative to neutral polymers brushes²⁵ and in other studies of adsorbed PE²⁹, arises when chains bridge the gap to adsorb simultaneously on both sliding surfaces. Bridging is absent for polymer brushes, which do not adsorb but are end-attached to the surface. The fact that, in contrast to all the other cases, the PE brushes uniquely retain their extreme propensity for lubrication even at volume fractions approaching unity, is compelling. It supports the notion that it is the counterion-induced PE-brush swelling, together with the fluid hydration sheaths about the charged segments and absence of bridging effects, that are effective in overcoming the factors leading to larger frictional dissipation for all other types of polymeric lubricants studied.

It is tempting to consider the relevance of this finding for other systems, particularly in nature, where charged polymers at biological surfaces are ubiquitous. In the experiments reported here water at very low salt concentration was used, but we note that

preliminary results with added salt (to concentration $\sim 10^{-2}$ M) show that the frictional forces remain extremely weak ($\mu_{\text{eff}} \lesssim 10^{-3}$), up to comparably high pressures. This salt concentration is still lower than the value (~ 0.1 M) pertaining in physiological conditions. However, the concentration of ionized monomers and the compensating free counterions within the PE brush can be easily comparable or larger than this when the brushes are highly compressed. This suggests that lubrication in living systems may be mediated by brush-like PE lubricants. □

Received 20 April; accepted 5 August 2003; doi:10.1038/nature01970.

1. Napper, D. H. *Polymeric Stabilization of Colloidal Dispersions* (Academic, London, 1983).
2. Dowson, D. (ed.) *Proc. Inst. Mech. Eng. H Vol. 201* (special issue on biolubrication) 189–247 (1987).
3. de Gennes, P. G. Polymers at an interface: A simplified view. *Adv. Colloid Interface Sci.* **27**, 189–207 (1987).
4. Israelachvili, J. N. *Intermolecular and Surface Forces* (Academic, London, 1992).
5. Klein, J., Kumacheva, E., Mahalu, D., Perahia, D. & Fetters, L. Reduction of frictional forces between solid surfaces bearing polymer brushes. *Nature* **370**, 634–636 (1994).
6. Grest, G. S. Normal and shear forces between polymer brushes. *Adv. Polym. Sci.* **138**, 149–182 (1999).
7. Kilbey, S. M. II, Schorr, P. A. & Tirrell, M. in *Dynamics of Small Confining Systems IV* (eds Drake, J. M., Grest, G. S., Klafter, J. & Kopelman, R.) 181–187 (Materials Research Society, Pittsburgh, PA, 1999).
8. Raviv, U. & Klein, J. Fluidity of bound hydration layers. *Science* **297**, 1540–1543 (2002).
9. Klein, J. & Kumacheva, E. Simple liquids confined to molecularly thin layers. I. Confinement-induced liquid to solid phase transitions. *J. Chem. Phys.* **108**, 6996–7009 (1998).
10. Pashley, R. M. Hydration forces between mica surfaces in aqueous electrolyte solutions. *J. Colloid Interface Sci.* **80**, 153–162 (1981).
11. Raviv, U., Laurat, P. & Klein, J. Fluidity of water confined to sub-nanometre films. *Nature* **413**, 51–54 (2001).
12. Tadmor, R., Rosensweig, R. E., Frey, J. & Klein, J. Resolving the puzzle of ferrofluid dispersants. *Langmuir* **16**, 9117–9120 (2000).
13. Raviv, U., Giasson, S., Frey, J. & Klein, J. Viscosity of ultra-thin water films confined between hydrophobic or hydrophilic surfaces. *J. Phys. Condens. Matter* **14**, 9275–9283 (2002).
14. Gohy, J. F., Antoun, S. & Jerome, R. Self-aggregation of PMMA-b-PSGMA copolymers. *Polymer* **42**, 8637–8645 (2001).
15. Abraham, T., Giasson, S., Gohy, J.-F., Jerome, R. & Stamm, M. Adsorption kinetics of a hydrophobic-hydrophilic diblock polyelectrolyte at the solid-aqueous solution interface. *Macromolecules* **33**, 6051–6059 (2000).
16. Witten, T., Leibler, L. & Pincus, P. Stress-relaxation in the lamellar copolymer mesophase. *Macromolecules* **23**, 824–829 (1990).
17. Wijmans, C. M., Zhulina, E. B. & Fleer, G. J. Effect of free polymer on the structure of a polymer brush and interaction between 2 polymer brushes. *Macromolecules* **27**, 3238–3248 (1994).
18. Klein, J. Shear, friction, and lubrication forces between polymer-bearing surfaces. *Ann. Rev. Mater. Sci.* **26**, 581–612 (1996).
19. Miklavic, S. J. & Marcelja, S. Interaction of surfaces carrying grafted polyelectrolytes. *J. Phys. Chem.* **92**, 6718–6724 (1988).
20. Pincus, P. Colloid stabilization with grafted polyelectrolytes. *Macromolecules* **24**, 2912–2919 (1991).
21. Misra, S., Tirrell, M. & Mattice, W. Interaction between polyelectrolyte brushes in poor solvents. *Macromolecules* **29**, 6056–6060 (1996).
22. Borisov, O. B., Zhulina, E. B. & Birshtein, T. M. Diagram of states of a grafted polyelectrolyte layer. *Macromolecules* **27**, 4795–4803 (1994).
23. Csajka, F. S., Netz, R. R., Seidel, C. & Joanny, J.-F. Collapse of polyelectrolyte brushes: Scaling theory and simulations. *Eur. Phys. J. E* **4**, 505–513 (2001).
24. Raviv, U. *et al.* Properties and interactions of physigrafted end-functionalised poly(ethylene glycol) layers. *Langmuir* **18**, 7482–7495 (2002).
25. Raviv, U., Tadmor, R. & Klein, J. Shear and frictional interactions between adsorbed polymer layers in a good solvent. *J. Phys. Chem. B* **105**, 8125–8134 (2001).
26. de Gennes, P. G. *Scaling Concepts in Polymer Physics* (Cornell Univ. Press, Ithaca, NY, 1979).
27. Ferry, J. D. *Viscoelastic Properties of Polymers*, 3rd edn (Wiley, New York, 1985).
28. Claesson, P. M. & Ninham, B. W. pH dependent interactions between adsorbed chitosan layers. *Langmuir* **8**, 1406–1412 (1992).
29. Ruths, M., Sukhishvili, S. A. & Granick, S. Static and dynamic forces between adsorbed polyelectrolyte layers (quaternized poly-4-vinylpyridine). *J. Phys. Chem. B* **105**, 6202–6210 (2001).
30. Derjaguin, B. V., Churaev, N. V. & Muller, V. M. *Surface Forces* (Plenum, New York, 1987).
31. Kampf, N., Raviv, U. & Klein, J. Normal and shear forces between adsorbed and gelled layers of chitosan, a naturally occurring cationic polyelectrolyte. *Macromolecules* (submitted).

Acknowledgements We thank J. Frey for the STAI synthesis, and R. Tadmor, T. Witten, D. Lukatsky and P. Pincus for discussions. We also thank the Eshkol Foundation (U.R.), the Canadian Friends of the Weizmann Institute (Chapman-Vered Fellowship, S.G.), the US-Israel BSF, the Deutsche-Israelische Program, and the Israel Science Foundation for their support of this work. J.F.G. (Chargé de Recherches by FNRS) and R.J. thank the SSTC for financial support.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.K. (jacob.klein@weizmann.ac.il or jacob.klein@chem.ox.ac.uk) or S.G. (suzanne.giasson@umontreal.ca)

No upward trends in the occurrence of extreme floods in central Europe

Manfred Mudelsee^{1*}, Michael Börngen¹, Gerd Tetzlaff¹
& Uwe Grünewald²

¹Institute of Meteorology, University of Leipzig, Stephanstrasse 3, 04103 Leipzig, Germany

²Institute of Hydrology, Technical University Cottbus, PF 101344, 03103 Cottbus, Germany

* Present address: Department of Earth Sciences, Boston University, 685 Commonwealth Avenue, Boston, Massachusetts 02215, USA

Extreme river floods have been a substantial natural hazard in Europe over the past centuries¹, and radiative effects of recent anthropogenic changes in atmospheric composition are expected to cause climate changes, especially enhancement of the hydrological cycle², leading to an increased flood risk^{3,4}. For the past few decades, however, observations from Europe^{1,5–7} do not show a clear increase in flood occurrence rate. Here we present longer-term records of winter and summer floods in two of the largest rivers in central Europe, the Elbe and Oder rivers. For the past 80 to 150 yr, we find a decrease in winter flood occurrence in both rivers, while summer floods show no trend, consistent with trends in extreme precipitation occurrence. The reduction in winter flood occurrence can partly be attributed to fewer events of strong freezing—following such events, breaking river ice at the end of the winter may function as a water barrier and enhance floods severely. Additionally, we detect significant long-term changes in flood occurrence rates in the sixteenth to nineteenth centuries, and conclude that reductions in river length, construction of reservoirs and deforestation have had minor effects on flood frequency.

The middle Elbe (between Litoměřice and Magdeburg) and middle Oder (between Racibórz and before Kostrzyn) drain basins (respectively 95,000 km² and 54,000 km²) that are under a continental, low-range mountainous climate (Sudeten Mountains, Bohemia, Erzgebirge and Beskids). Floods in hydrological summer (May to October) are caused by heavy rainfall, and in the winter also by thawing snow^{8,9}. Breaking river ice may function as a barrier, enhancing winter floods severely⁹.

Weikinn's sources¹⁰ contain 23,160 documentary entries on hydrographic events in Europe up to 1850, with high coverage of Germany and neighbouring countries. Entries were checked and augmented using further sources (see Supplementary Information) on historical Elbe and Oder floods. Local events were excluded by requiring information¹⁰ supportive of a regional extent, namely: (1) floods at other places on the river, (2) floods of tributaries at the same time, and (3) favourable meteorological conditions such as large rainfall before a flood. Stage values, available for Elbe floods at Dresden, Pillnitz and Meißen^{11–13} and for the Oder at Krosno and other stations^{8,14}, facilitated construction of an impact-related magnitude scale. We restricted the number of magnitude classes to three (1, minor; 2, strong; 3, exceptionally strong), as in refs 1 and 15. This reduction of information is balanced by a gain in robustness of the constructed records: errors in judging the magnitude of a flood (by witnesses, source authors, Weikinn or us), have a minor influence when class bounds are wide. Likewise, uncertainties in the stage–runoff relations (see below) are less important than when more classes were used.

Although Weikinn aimed to compile original sources by witnesses, he also listed other, unverified references, and did not use historical source interpretation. Therefore, to assess the reliability of the constructed flood records (Supplementary Information), a comparison was made with CLIMDAT¹⁶, a climate database of

historically and critically reviewed documents. CLIMDAT focuses on central and eastern Germany, Poland and the Czech Republic, and covers the interval AD 1500 to 1799. For class-3 events, the agreement ("Weikinn's" floods (year, season, month) listed in CLIMDAT) is 100% (Elbe, Oder), for class-2 events 72% (Elbe) and 65% (Oder), and for class-1 events 39% (Elbe) and 29% (Oder). Floods listed in CLIMDAT but not by Weikinn are few (Elbe, 1.5%; Oder, 10%) and probably of minor magnitude. Thus, the information provided by Weikinn on heavy floods (classes 2–3) after ~1500 probably has few errors (for example, duplication of events caused by misrecording). This is confirmed by the agreement between flood occurrence rates (see below) calculated from the Weikinn sources and CLIMDAT, which are indistinguishable within the confidence bands. On the other hand, our records may not be homogenous before ~1500 and for class-1 floods. Therefore we restrict statistical analyses to heavy floods after 1500. This conservative approach does not imply that the class-1 records are useless; Weikinn may have had access to information that has now been lost.

The Elbe flood record was extended to the present using measurements from Dresden: monthly maximum water stage from 1850 to 1892 (ref. 17), and daily runoff from 1852 to 2002 (Global Runoff Data Centre (GRDC), <http://www.wetteronline.de>). The Dresden runoff time series exhibits excellent correlations with daily runoff series from Elbe stations Děčín (1887–1990, lag = 0, $r^2 = 0.98$) and Barby (1899–1999, lag = 2 days, $r^2 = 0.91$), which underline the regional character of the constructed record. Note that daily runoff was rarely measured directly but rather inferred via stage by frequently updated stage–runoff relations, introducing uncertainty. For Dresden this relation was found¹⁸ to have changed only slowly over time, and to be valid also for upper values. For the interval 1852–92, this assessment is supported (Fig. 1a) by the high correlation ($r^2 = 0.90$) and a fit uncertainty that is clearly smaller than the class widths. Accuracy of runoff values, record length and availability of pre-1850 values^{11–13} make Dresden the most suitable station to study middle Elbe floods.

The Oder record was extended using monthly maximum stage from Krosno from 1905 to 1936 (ref. 19), and daily runoff from Eisenhüttenstadt (40 km downstream of Krosno) from 1920 to 2002 (GRDC, <http://www.wetteronline.de>). Although the correlation in Fig. 1b is high ($r^2 = 0.91$), our constructed Oder record may miss some minor floods between 1850 and 1904 because for that interval only flood measurements^{8,14}, not continuous data, are available. The correlation between Eisenhüttenstadt and Polecko (monthly runoff from 1946 to 1987) is good ($r^2 = 0.84$). Precise information on

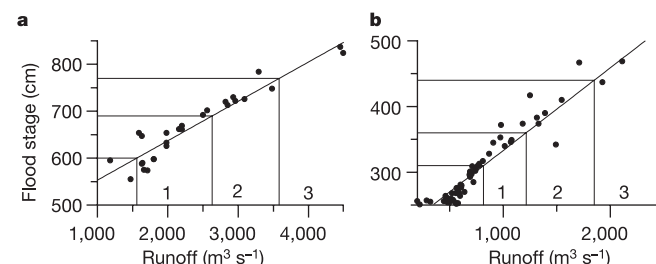


Figure 1 Classification of flood magnitudes (1: minor, 2: strong, 3: exceptionally strong) using linear regressions between flood stage and inferred runoff. Thresholds were applied to documented floods (before 1850) and measurements (after 1850). **a**, Elbe, station Dresden, 1852–92, $n = 27$, $r^2 = 0.90$, $\sigma_s = 24$ cm, $\sigma_r = 286$ m³ s^{−1}, thresholds: 600/690/770 cm and 1,560/2,630/3,580 m³ s^{−1}. **b**, Oder, stations Krosno (stage) and Eisenhüttenstadt (runoff), 1891–1936, $n = 58$, $r^2 = 0.91$, $\sigma_s = 18$ cm, $\sigma_r = 143$ m³ s^{−1}, thresholds: 310/360/440 cm and 815/1,215/1,850 m³ s^{−1}. n , data size; r , correlation coefficient; σ_s (fit uncertainty in stage) = $(\chi^2/(n - 2))^{1/2}$; χ^2 , sum of squares; σ_r (fit uncertainty in runoff) = σ_s/slope . In **a** and **b**, the widths of magnitude classes are about 3–4 times larger than the fit uncertainties.

stage–runoff relation seems not to be available⁹ owing to war chaos and divided responsibilities. For the interval 1891–1936 the high correlation/low fit uncertainty (Fig. 1b), however, indicates stability of the stage–runoff relation at upper values. In terms of record length, stability of stage–runoff relation and availability of pre-1920 values^{8,14,19}, Krosno/Eisenhüttenstadt is suited to analyse floods of the middle Oder, although not of the same quality as Dresden for the Elbe.

Elbe and Oder floods (magnitudes 1–3) show a clear seasonal dependence, with winter floods occurring mainly in February–March (Elbe, Oder) and summer floods in June–July (Elbe) or August (Oder). (For the Elbe from 1021 to 2002, the number of floods in January/February/.../December is 51/76/81/35/18/38/35/25/8/4/6/27; for the Oder from 1269 to 2002, the number of floods is 19/24/50/30/25/23/27/34/15/9/3/11.) Before 1850, of 103 Elbe winter floods that allowed unambiguous distinction, 91 were connected with a frozen river (documented in the Weikinn/CLIMDAT sources); for the Oder, the number is 28 out of 34 events. Data (Supplementary Information) show further that during 1930–70, only 2 out of 13 Elbe winter floods were influenced by freezing; for the Oder, the number is 3 out of 20; in both rivers, the last ice flood occurred in 1947. For summer floods at Dresden and Eisenhüttenstadt, the 100-yr runoff value, estimated by fitting a generalized extreme value (GEV) distribution to bi-monthly maxima²⁰, is $3,900 \text{ m}^3 \text{ s}^{-1}$ and $2,500 \text{ m}^3 \text{ s}^{-1}$, respectively. The correlation coefficients (calculated separately for winter and summer) between Elbe and Oder flood time series (constructed using monthly magnitude data with magnitudes for months without a flood set equal to zero) during 1500–2002 are around 0.3. Analyses of extracted intervals confirmed that these values have not changed over time. The weak relation between individual Elbe and Oder floods shows that orographic differences between both catchment areas are effective. This is illustrated by the following examples of major floods (see Supplementary Information): August 1501 (Elbe, Oder), February–March 1784 (Elbe), April 1785 (Elbe, Oder), July 1997 (Oder), August 2002 (Elbe).

During the interval ~1820–2002, the estimated occurrence

rate^{21,22} (see Methods) of winter floods of the Elbe steadily decreased, while that of summer floods did not change significantly (Fig. 2). The statistical test (see Methods), using measured runoff (Q) data from Dresden (1852–2002), confirmed these results at the 90% level—not only for the magnitude classes that we used ($Q \geq 1,560/2,630 \text{ m}^3 \text{ s}^{-1}$), but also for other thresholds ($1,450/1,800/2,200/3,000 \text{ m}^3 \text{ s}^{-1}$), attesting to the robustness of the trends. Using data from stations Děčín and Barby gave similar test results. In the Oder during ~1800–1920, occurrences of winter as well as summer floods increased (Fig. 2). This should, however, be interpreted cautiously, as data quality in 1850–1904 is reduced. For 1920–2002, the Oder exhibited the same trends as the Elbe: less winter floods and no change in summer flood occurrence, confirmed by the test for various thresholds ($Q \geq 600/700/815/1,110/1,215/1,350/1,500 \text{ m}^3 \text{ s}^{-1}$). (Using monthly runoff data from Oder station Polecko did not reveal significant trends owing to the shortness of the record.) Attempts to identify short-term changes (last ~30 yr) or changes in class-3 flood occurrence lacked statistical power.

To evaluate whether uncertainties in the stage–runoff relation (Fig. 1) are small enough not to corrupt trend test results, a simulation study was carried out (see Methods). This showed that adding noise up to twice the uncertainty ($2\sigma_r$) to the runoff time series left results unchanged, making the trends robust also in that respect.

To test whether the observed trends over the past ~80–150 yr were the effect of increases^{9,23} (since ~1900) in total reservoir size, we constructed flood records (Supplementary Information) that correct for this effect (see Methods) and repeated the trend test. In case of heavy floods (classes 2–3), the results were the same as for the uncorrected records (Fig. 2 and Supplementary Table). This means that for the rivers Elbe and Oder, reservoir sizes are too small to influence the occurrence of heavy floods. On the other hand, the occurrence of minor floods (class 1) can be affected (Supplementary Table). Caution is required as regards the upward trend of reservoir-size-corrected summer floods (classes 1–3) of the Elbe. Because in practice total reservoir size cannot be used to 100%, the test

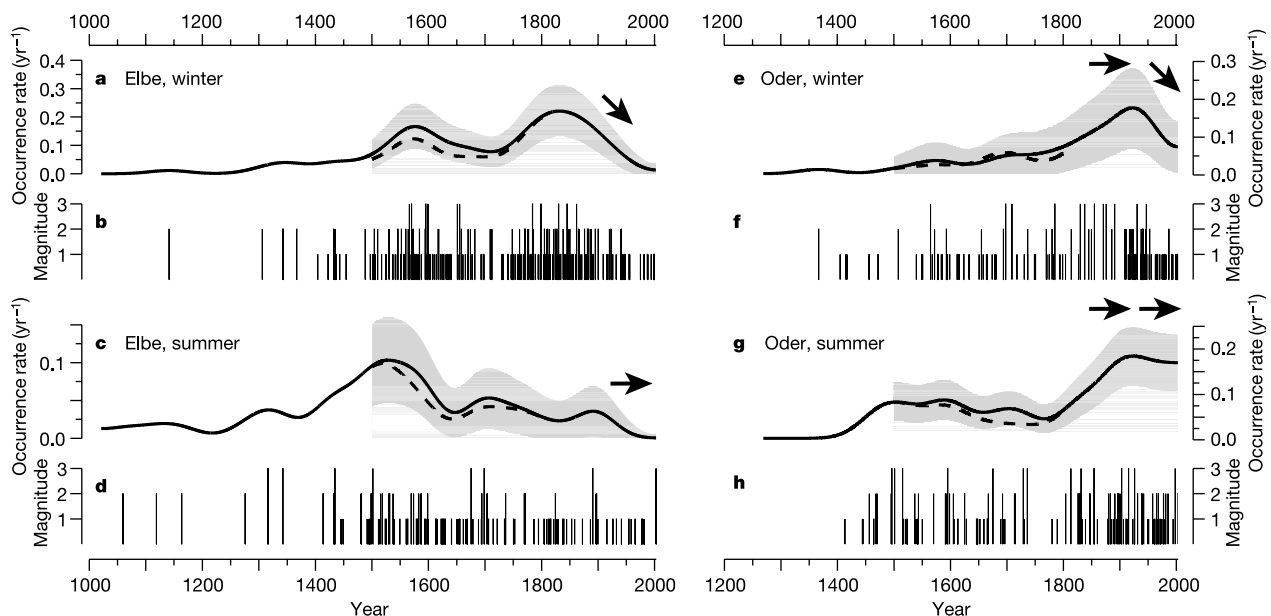


Figure 2 Occurrence rates of heavy floods (magnitude classes 2–3) in central Europe. **a, b**, Elbe, winter; **c, d**, Elbe, summer; **e, f**, Oder, winter; **g, h**, Oder, summer. Flood data from Weikinn's sources¹⁰ (Supplementary Information) (**b, d, f, h**) were analysed using a gaussian kernel, a bandwidth of 35 yr and bootstrap simulations (see Methods). This yielded (**a, c, e, g**) occurrence rates (solid lines) and 90% confidence bands (grey); occurrence rates using data from CLIMDAT¹⁶ (Supplementary Information) for 1500–

1799 are shown as dashed lines. Records before 1500 are probably not homogenous (no confidence bands drawn). Arrows indicate the results (downward/no trend) from the statistical test (90% level) for trend in the flood occurrence rate (Elbe, 1852–2002; Oder, 1850–1920 and 1920–2002); results for uncorrected and reservoir-size-corrected data are identical. For occurrence rates and trends including data on minor floods (class 1), see Supplementary Information.

becomes liberal for upward trends (see Methods), and the apparent upward trend may be an artefact. Because land-use changes^{9,18} over the past ~80–150 yr (for example, deforestation in the mountainous catchment parts) would, if effective, reduce retention capability²⁴ and induce upward trends (which are absent for heavy floods), we assess their influence as negligible.

To assess whether climate variations had an influence on detected downward trends of winter flood occurrence and on absent trends of summer flood occurrence, we analysed records of precipitation in central Europe. Data²⁵ consist of homogenized monthly estimates from 1900 to 1999 for gridboxes at 2.5° latitude by 3.75° longitude resolution. The gridbox centred at 50° N, 15° E contains nearly the entire catchment area of the Elbe at Dresden and parts of the catchment area of the Oder at Eisenhüttenstadt; the gridbox at 50° N, 18.75° E contains eastern and southern parts of the Oder catchment, but mainly other areas. Winter (November–April) and summer precipitation was treated separately. Statistically significant trends in the occurrence of 25-yr maxima were detected (Fig. 3). Comparing these trends with those of extreme floods (Fig. 2) over the past ~100 yr roughly indicates a causal influence of extreme precipitation events on floods. However, besides parallel trends (winter floods/precipitation at 50° N, 15° E; summer floods/precipitation at 50° N, 18.75° E) we find also non-parallel (but not opposing) trends (Figs 2, 3). We explain this discrepancy as follows. (1) Limited temporal resolution (floods, precipitation) and (2) limited spatial resolution (precipitation) may blur a causal connection. (3) Runoff-influencing factors are having an effect. Such factors presumably consist less of reservoir-size or land-use changes (see above) and more of climate changes. In the case of winter floods, a factor could be the reduced rate of strong river freezing (that is, potentially flood enhancing) events (see above and ref. 1). Reduced freezing may have been caused by warming²⁶ or increased pollution of river waters. Elevated winter temperatures could also have had an effect via a reduction of occurrence of frozen soil, which has low absorbing capacity²⁴.

Figure 2 indicates further that the flood occurrence rates increased markedly during the fourteenth and fifteenth centuries. These increases are probably an artefact caused by inhomogeneous records: fewer documents about floods from that period—compared with following periods—were written (before the invention of printing) or have lasted to the present day, or flood perception was different¹, leading to missed events. A maximum in flooding rate was reached in the sixteenth century (presumably in its latter half), and was more pronounced in case of the Elbe (Fig. 2). Rivers in central and southwest Europe during the sixteenth century had a

similar increase¹⁵, which was attributed to higher precipitation. The subsequent reduction in Elbe winter flood occurrence to a low around 1700 (Fig. 2a) might reflect the cold and dry climate of the Late Maunder Minimum²⁷. However, the Oder (Fig. 2e) shows no such reduction. Major reductions in length²⁸ of the middle Elbe were made during 1740–1870, and reductions in length of the middle Oder were made during 1745–1850. A consistent (winter/summer, Elbe/Oder) signal of these length reductions is not seen in the flood data (Fig. 2). This suggests a minor influence of length reductions on the occurrence of heavy floods—a detailed hydrological model would be required to quantify that influence.

Although extreme floods with return periods of 100 yr and more occurred in central Europe in July 1997 (Oder) and August 2002 (Elbe), there is no evidence from the observations for recent upward trends in their occurrence rate. Global climate changes affect many and various processes in regional hydrology, such as river and soil freezing in the case of winter floods under continental climate. Long, continuous, seasonally and regionally resolved flood records like those for the Elbe and Oder should facilitate evaluation of future trends using coupled atmospheric–hydrological models²⁴. □

Methods

Occurrence rate estimation

Kernel estimation^{21,22} allows detailed inspection of time-dependent flood occurrence rates and assessment of significant changes with the help of confidence bands. We used a gaussian kernel, K , to weigh observed flood dates, $T(i)$, $i = 1, \dots, N$ (number of floods), and calculate the occurrence rate, λ , at time t as:

$$\lambda(t) = \sum_i K((t - T(i))/h) \quad (1)$$

Selection of bandwidth ($h = 35$ yr) was guided by cross-validation²¹. Confidence bands (90%) around $\lambda(t)$ were determined using a bootstrap technique: N simulated floods were drawn from $T(i)$ with replacement and simulated λ calculated. This procedure was repeated 2,000 times, and a percentile- t confidence band²¹ calculated. Detected trends in occurrence rate were confirmed for the measured interval (1850–2002) using the statistical test in ref. 29. Under the null hypothesis H_0 'constant occurrence rate', the quantity $[\sum_i T(i)/N - (t_u + t_l)/2] / [(t_u - t_l)/(12N)^{1/2}]$ is standard-normally distributed. Here t_u is the upper bound of the observation interval (September 2002), t_l the lower bound (1850). H_0 was tested against one-sided alternatives (for example, ' λ increases') at the 90% level. Robustness of test results against uncertainties in stage–runoff relations was confirmed as follows. For each station, 2,000 simulated runoff time series were generated by adding noise (2σ , which means a conservative approach) to measured data; simulated flood dates were determined and trend tests repeated. This yielded the same average trends.

Reservoir-size correction

The time-dependent total volume, $V(t)$, of reservoirs above a station that can be employed for flood management was used to construct corrected flood records that assume a constant reservoir size at present level, V_p . $V(t)$ increased monotonically since ~1900; data^{9,23} are given in Supplementary Information. To calculate reduced runoff, Q_r , for a flood that could have been obtained if $V(t)$ equalled V_p , we 'cut off' the flood peaks in daily runoff, $Q(t)$:

$$\int (Q(t) - Q_r) dt = V_p - V(t) \quad (2)$$

where the integral is over the time a flood lasted. Q_r was then sorted into the same magnitude classes as in Fig. 1. This was carried out for all detected floods (Elbe, Dresden; Oder, Eisenhüttenstadt) in the instrumental period, yielding records (Supplementary Information) that are not influenced by changes in reservoir size. Note that in practice utilization of the total reservoir size is less than 100%, making the Q_r estimate a lower bound. This further implies that trend tests using reservoir-size-corrected records become conservative (liberal) as regards downward (upward) trends.

Received 16 May; accepted 16 July 2003; doi:10.1038/nature01928.

- Glaser, R. *Klimageschichte Mitteleuropas* (WBG, Darmstadt, 2001).
- Giorgi, F. et al. in *Climate Change 2001: The Scientific Basis* (eds Houghton, J. T. et al.) 583–638 (Cambridge Univ. Press, Cambridge, 2001).
- Palmer, T. N. & Räisänen, J. Quantifying the risk of extreme seasonal precipitation events in a changing climate. *Nature* **415**, 512–514 (2002).
- Christensen, J. H. & Christensen, O. B. Severe summertime flooding in Europe. *Nature* **421**, 805–806 (2003).
- Disse, M. & Engel, H. Flood events in the Rhine basin: Genesis, influences and mitigation. *Nat. Hazards* **23**, 271–290 (2001).
- Sturm, K. et al. Hochwasser in Mitteleuropa seit 1500 und ihre Beziehung zur atmosphärischen Zirkulation. *Petermanns Geogr. Mitt.* **145**, 14–23 (2001).
- Robson, A. J. Evidence for trends in UK flooding. *Phil. Trans. R. Soc. Lond. A* **360**, 1327–1343 (2002).
- Fischer, K. Die Sommerhochwasser der Oder von 1813 bis 1903. *Jb. Gewässer. Norddeutsch. Besond. Mitt.* **1**, 1–101 (1907).

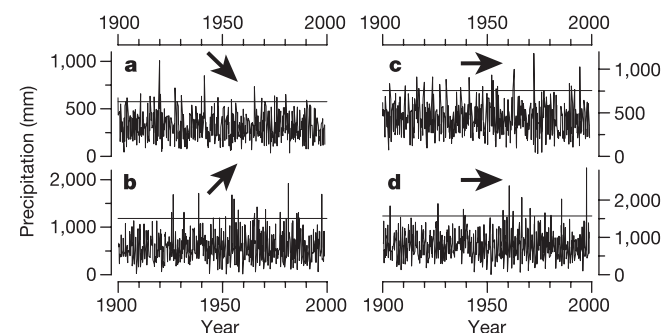


Figure 3 Occurrence of extreme precipitation events in central Europe during the twentieth century. Data are gridded monthly precipitation values²⁵ at 50° N, 15° E (a, b) and 50° N, 18.75° E (c, d) during winter (November–April) (a, c) and summer (May–October) (b, d). Shown as horizontal lines are 25-yr maxima (a, 576 mm; b, 1,182 mm; c, 755 mm; d, 1,573 mm), obtained by fitting a GEV distribution²⁰. Arrows indicate the results (upward/downward/no trend) from the statistical test (90% level) for a trend in the occurrence of 25-yr maxima.

9. Grünwald, U. *et al.* Ursachen, Verlauf und Folgen des Sommer-Hochwassers 1997 an der Oder sowie Aussagen zu bestehenden Risikopotentialen, Langfassung (IDNDR, Bonn, 1998).
10. Weikinn, C. Quellentexte zur Witterungsgeschichte Europas von der Zeitwende bis zum Jahre 1850: Hydrographie, Parts 1–4 (Akademie, Berlin, 1958–1963); Parts 5–6 (eds Börngen, M. & Tetzlaff, G.) (Gebrüder Borntraeger, Berlin, 2000–2002).
11. Pötzsch, C. G. Chronologische Geschichte der großen Wasserfluthen des Elbstroms seit tausend und mehr Jahren (Walthersche Hofbuchhandlung, Dresden, 1784).
12. Schäfer, W. Chronik der Dresdener Elbbrücke, nebst den Annalen der größten Elbfluthen von der frühesten bis auf die neueste Zeit (Adler und Dietze, Dresden, 1848).
13. Königliche Elbstrombauverwaltung Der Elbstrom, sein Stromgebiet und seine wichtigsten Nebenflüsse Vol. 3.1 (Reimer, Berlin, 1898).
14. Bureau des Ausschusses zur Untersuchung der Wasserverhältnisse in den der Überschwemmungsgefahr besonders ausgesetzten Flußgebieten Der Oderstrom, sein Stromgebiet und seine wichtigsten Nebenflüsse Vol. 3.1 (Reimer, Berlin, 1896).
15. Brázdil, R. *et al.* Flood events of selected European rivers in the sixteenth century. *Clim. Change* **43**, 239–285 (1999).
16. Miltzer, S. Klima, Umwelt, Mensch (1500–1800): Studien und Quellen zur Bedeutung von Klima und Witterung in der vorindustriellen Gesellschaft (Historisches Seminar, Leipzig, 1998).
17. Königliche Elbstrom-Bauverwaltung Hydrologischer Jahresbericht von der Elbe für 1892 (Baensch, Magdeburg, 1893).
18. Helms, M., Ihringer, J. & Nestmann, F. in *Morphodynamik der Elbe* (eds Nestmann, F. & Büchele, B.) 91–202 (Institut für Wasserwirtschaft und Kulturtechnik TH Karlsruhe, Karlsruhe, 2002).
19. Oderstrombauverwaltung Mitteilungen der Oderstrombauverwaltung über die Strombauten, die Wasserstände und die Schifffahrtsverhältnisse der Oder, 1905–1906, 1911–1926, 1927–1929, 1930–1932, 1933–1936 (Oderstrombauverwaltung, Breslau, 1907–36).
20. Reiss, R.-D. & Thomas, M. *Statistical Analysis of Extreme Values* (Birkhäuser, Basel, 1997).
21. Cowling, A., Hall, P. & Phillips, M. J. Bootstrap confidence regions for the intensity of a Poisson point process. *J. Am. Statist. Assoc.* **91**, 1516–1524 (1996).
22. Mudelsee, M. XTREND: A computer program for estimating trends in the occurrence rate of extreme weather and climate events. *Sci. Rep. Inst. Meteorol. Univ. Leipzig* **26**, 149–195 (2002).
23. Bundesanstalt für Gewässerkunde, *Untersuchungen zum Abflussregime der Elbe*, (BfG, Berlin, 2000).
24. Bronstert, A., Bismuth, C. & Menzel, L. (eds) *Proc. European Conf. on Advances in Flood Research* (Report No. 65, Potsdam Institute for Climate Impact Research, Potsdam, 2000).
25. Hulme, M., Osborn, T. J. & Johns, T. C. Precipitation sensitivity to global warming: Comparison of observations with HadCM2 simulations. *Geophys. Res. Lett.* **25**, 3379–3382 (1998).
26. Folland, C. K. *et al.* in *Climate Change 2001: The Scientific Basis* (eds Houghton, J. T. *et al.*) 99–181 (Cambridge Univ. Press, Cambridge, 2001).
27. Luterbacher, J. *et al.* The late Maunder Minimum (1675–1715)—A key period for studying decadal scale climatic change in Europe. *Clim. Change* **49**, 441–462 (2001).
28. Schmidt, M. *Hochwasser und Hochwasserschutz in Deutschland vor 1850* (Oldenbourg, Munich, 2000).
29. Cox, D. R. & Lewis, P. A. W. *The Statistical Analysis of Series of Events* (Methuen, London, 1966).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank J. Jacobeit, J. Luterbacher and M. Schulz for comments, and N. Conrads, H. Engel, W. Fröhlich, T. Lüllwitz, J. Munzar, R. Oppermann as well as GRDC (Germany) and NOAA (USA) for data. This work was supported by the Deutsche Forschungsgemeinschaft.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.M. (mudelsee@uni-leipzig.de).

Radiometric dating of the Siloam Tunnel, Jerusalem

Amos Frumkin¹, Aryeh Shimron² & Jeff Rosenbaum³

¹Department of Geography, The Hebrew University of Jerusalem, 91905, Israel

²Geological Survey of Israel, 30 Malkhei Israel Street, Jerusalem, 95501, Israel

³Postgraduate Research Institute for Sedimentology, Reading University, Reading RG6 6AB, UK

The historical credibility of texts from the Bible is often debated when compared with Iron Age archaeological finds (refs. 1, 2 and references therein). Modern scientific methods may, in principle, be used to independently date structures that seem to be mentioned in the biblical text, to evaluate its historical authenticity. In reality, however, this approach is extremely difficult because of poor archaeological preservation, uncertainty in identification,

scarcity of datable materials, and restricted scientific access into well-identified worship sites. Because of these problems, no well-identified Biblical structure has been radiometrically dated until now. Here we report radiocarbon and U–Th dating of the Siloam Tunnel^{3–10}, proving its Iron Age II date; we conclude that the Biblical text presents an accurate historic record of the Siloam Tunnel's construction. Being one of the longest ancient water tunnels lacking intermediate shafts^{11,12}, dating the Siloam Tunnel is a key to determining where and when this technological breakthrough took place. Siloam Tunnel dating also refutes a claim¹³ that the tunnel was constructed in the second century BC.

The Siloam Tunnel carries water into ancient Jerusalem from the only perennial spring of the region, the Gihon (Fig. 1). This tunnel has attracted the attention of many scholars, who have yet to solve some of its mysteries^{3–13}. Most scholars ascribe the Siloam Tunnel to King Hezekiah (727–698 BC), following the biblical verses (2 Kings 20:20, 2 Chronicles 32:3,4) narrating that this Judahite king constructed a waterwork that “brought water into the city”. In addition to the biblical record, the Siloam Tunnel's association with Hezekiah also relies on the palaeography and philology of the monumental Siloam inscription found close to the Siloam Tunnel's outlet (refs 14, 15 and references therein). However, the name of King Hezekiah is not mentioned in the inscription, unlike other monumental inscriptions of the Levant, which often praised monarchs for their architectural achievements. If the assumed date of the inscription is correct, it can still be argued that the inscription could be a later copy of a literary narrative, such as the ‘Chronicles of the Kings of Judah’, so the Siloam Tunnel itself may be older than the inscription. Such a view could be supported by the unique style, differing from most other North-Semitic inscriptions¹⁵. According to ref. 13, a

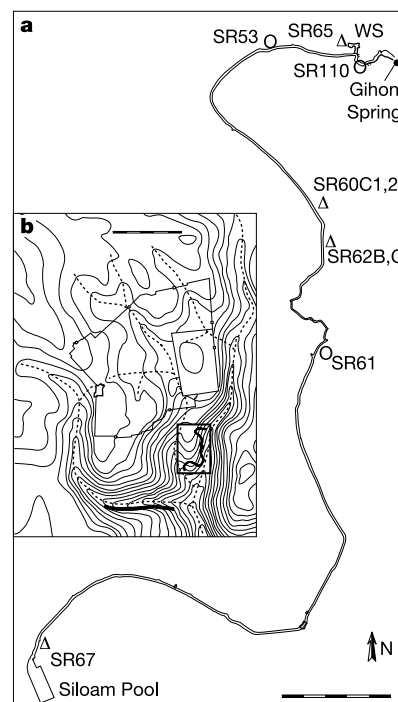


Figure 1 Location map. **a**, Plan of the Siloam Tunnel (Long. 35.17°; Lat. 31.68°) with the location of dated samples. Scale bar, 50 m. Circle, plant fragments within plaster, dated by ¹⁴C AMS; triangle, speleothem sample dated by ²³⁰Th–²³⁴U thermal ionization mass spectrometry. WS, Warren Shaft. **b**, Relation of the Siloam Tunnel (bold line) to the topography of Jerusalem and the present city walls. The Siloam Tunnel conveyed the water of the Gihon Spring into the Siloam Pool within the original core of Jerusalem. Contours are 10 m (vertically) apart. Scale bar, 500 m.

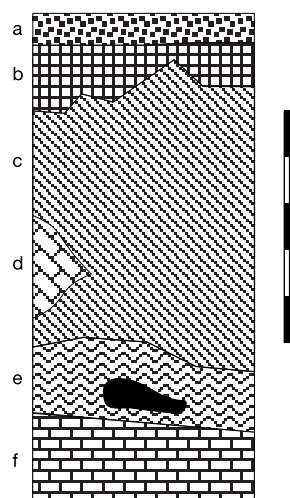


Figure 2 Typical layering in a core drilled at the bottom of the Siloam Tunnel. Scale bar, 5 cm. Layer a, silty tufa; layer b, young black plaster; layer c, intermediate coarse plaster; d, limestone clast; layer e, the lowest plaster termed 'ancient plaster', in which a piece of wood is shown (black); layer f, limestone bedrock. Note that no deposit exists between the ancient plaster and the bedrock.

much later date should be assigned to the inscription and the Siloam Tunnel: about the second century BC, and the Siloam inscription cannot serve as a cornerstone for dating other inscriptions, because it was never dated independently. "Attempts to reconstruct the history of Hebrew scripts can only be based upon texts whose dating can be established other than by palaeography"¹³. Other scholars¹⁶ discredited the proposed late age of the Siloam Tunnel, but both sides of the debate relied on the inscription text itself, failing to produce any independent dating. Most scholars believe that independent dating of the Siloam Tunnel is impossible because the tunnel was cleared of debris during the early twentieth century³, so it does not contain any artefacts that can be dated by classic archaeological methods.

We overcome this difficulty by dating artificial and natural materials incorporated within the tunnel walls, ceiling and floor, which constrain the age of the Siloam Tunnel. During our geological mapping of the tunnel some of our work focused on an important feature (previously documented in ref. 3) that the Siloam Tunnel was originally cemented with plaster covering its floor, walls and occasionally ceiling. This material had been applied to the rock-cut limestone and dolomite surface of the Siloam Tunnel to seal it and to avoid the loss of water through fissures. Such fissures in the bedrock above and within the Siloam Tunnel allow dripwater to

enter the tunnel and deposit calcite speleothems (flowstone and stalactites). We have drilled and obtained core samples from these materials, and classified them according to field relations as well as petrographic, geochemical, scanning electron microscopy (SEM), X-ray diffraction (XRD) and isotopic analyses, performed at the Geological Survey of Israel laboratories¹⁷.

The plaster of the Siloam Tunnel varies in thickness from about 1 to 20 cm. In a typical core we distinguish a silty tufa veneer covering three varieties of plaster (Fig. 2). We refer to the stratigraphically oldest variety as the 'ancient plaster' (AP)¹⁷. The ancient plaster is a very fine hydraulic lime plaster, composed of a recarbonated lime (CaO) binder with a filler of soil, chips of marl, organic materials and bone fragments, as well as rare charcoal and ash. Shrinkage cracks and oval vugs are abundant and are now filled with acicular crystals of aragonite. XRD and SEM also reveal the presence of bone apatite, traces of the rare carbonate vaterite and zoned isotropic peloids containing amorphous varieties of CaAl-silicates with fibres of ettringite. The low firing temperature, absence of pottery shards and the relatively high content of clays and phosphorus clearly distinguish the ancient plaster from younger plasters in the waterworks. The chemical and petrographic characteristics of the ancient plaster indicate that the source rock for lime manufacture was probably the phosphorus-rich marly chalk occurring in Senonian formations on the hills east of Jerusalem. The marly and organic-rich nature of this chalk imparted upon the plaster its natural hydraulic characteristics. This plaster is generally covered by younger plasters or by a lamina, about 2–4 cm thick, of silty tufa and siltstone deposited from the spring waters flowing along the Siloam Tunnel. The siltstone is quartzose with some feldspar and abundant heavy minerals (magnetite, ilmenite, amphibole, garnet and zircon). It is derived by the erosion of a sandy unit intercalated with the local carbonate rock.

We drilled 15 holes through the Siloam Tunnel's floor, evenly distributed along the entire length of the tunnel. None of them encountered evidence of a sedimentary layer between the ancient plaster and bedrock. Because clastic and chemical sediments were, and are to this day, deposited from waters of the spring, their absence beneath the ancient plaster indicates that a natural water conduit was not present along the bottom of the Siloam Tunnel before the application of the ancient plaster. It also indicates that plastering took place soon after construction of the tunnel. Plastering of ancient waterworks has been a well-known technological practice in the region^{12,18}.

We have recovered from the ancient plaster fine plant fragments that are extraordinarily well preserved. Two samples were radiocarbon dated at the Oxford University accelerator mass spectrometry (AMS) facility by its standard procedure¹⁹. After correction for isotopic fractionation with the use of measured $\delta^{13}\text{C}$, and based on the ^{14}C half-life of 5,568 yr, the conventional dates for the ancient plaster organic materials are $2,620 \pm 35$ yr BP

Table 1 Radiogenic isotopic data and ages of Siloam Tunnel (ST) and the bottom of Warren Shaft (WS)

Field no.	Laboratory no.	Sample description	$\delta^{13}\text{C}$	^{14}C date (yr BP)	Calibrated age, 1 σ range
SR53*	OxA-8522	Piece of wood, ST	-26.0	2,620 $\pm$ 35	822-796 BC
SR61*	OxA-8523	Short-lived plant, ST	-25.0	2,505 $\pm$ 35	790-760; 690-540 BC
SR110*	OxA-8990	Piece of wood, ST	-27.5	605 $\pm$ 65	AD 1300-1405

Field no.	Sample description	U conc. (p.p.m.)	$^{230}\text{Th}/^{234}\text{U}$	$^{234}\text{U}/^{238}\text{U}$	$^{230}\text{Th}/^{232}\text{Th}$	$^{230}\text{Th}/^{234}\text{U}$ age (yr)
SR67†	Stalactite from ceiling, ST	0.617	0.02237 $\pm$ 0.00016	1.0403 $\pm$ 0.0031	27.15	2,317 $\pm$ 18
SR65†	Flowstone in natural void, WS	0.386	0.9899 $\pm$ 0.0054	1.0918 $\pm$ 0.0052	920.5	>400,000
SR62B†	Flowstone on plaster, ST	0.911	0.01942 $\pm$ 0.00020	1.2253 $\pm$ 0.0053	2.446	<1,000
SR62C†	Flowstone on plaster, ST	0.615	0.05892 $\pm$ 0.00056	1.2185 $\pm$ 0.0035	1.604	<500
SR60C1†	Flowstone on plaster, ST	0.664	0.05082 $\pm$ 0.00083	1.3765 $\pm$ 0.0027	2.619	<2,500
SR60C2†	Flowstone on plaster, ST	0.625	0.03101 $\pm$ 0.00014	1.3956 $\pm$ 0.0048	2.067	<2,000

*Radiocarbon AMS dates calibrated with OxCal^{20,21}.

† ^{230}Th – ^{234}U dates of speleothems corrected for detrital content assuming that initial $^{230}\text{Th}/^{232}\text{Th} = 1.6$ (ref. 22).

Isotope ratios are given as activity ratios.

for a piece of wood (SR53; Table 1), and $2,505 \pm 35$ yr BP for a short-lived plant (SR61). Calibrated with the OxCal program²⁰ using the INTCAL 98 calibration curve²¹, these correspond within 1σ uncertainty to calendar age ranges of 822–796 BC for the wood sample, and a multiple range of 790–760 and 690–540 BC for the short-lived plant. Within an uncertainty margin of 2σ , the short-lived plant age range is 800–510 BC.

To test these dates against the observed stratigraphic order of plasters, we also dated a small fragment of charred wood recovered from a layer of coarse slag-bearing plaster overlying the ancient plaster between the Gihon Spring and Warren Shaft. The conventional ^{14}C age of this sample (SR110; Table 1) is 605 ± 65 yr BP, which is in agreement with the field relations and indicates previously undocumented Mamluk period (AD 1260–1516) activity in the Siloam Tunnel.

The chronology of the Siloam Tunnel speleothems is based on ^{230}Th – ^{234}U dating by thermal ionization mass spectrometry performed at the Open University, Milton Keynes (Table 1). Laboratory errors are reported with 2σ uncertainty. After conventional chemical preparation, mass abundances of natural U and Th isotopes were measured²². For precise dating we used only reliable samples in which $^{230}\text{Th}/^{232}\text{Th} > 20$, indicating that radiogenic ^{230}Th predominated and detrital contamination was negligible²³. We calculated the dates and corrected for detrital content assuming an initial $^{230}\text{Th}/^{232}\text{Th}$ ratio of 1.6 (ref. 22).

Stalactite SR67, taken from the ceiling of the Siloam Tunnel about 4 m above the flowing water, was sufficiently distal from detrital contamination, as also indicated by its low $^{230}\text{Th}/^{232}\text{Th}$ ratio. The stalactite is not active today, so its deposition must have occurred closer to the Siloam Tunnel's construction than to the present. Its thorium age (measured on AD 2000) is $2,317 \pm 18$ yr, representing the mean deposition age of the bulk sample. This indicates that the Siloam Tunnel was constructed before 2,300 yr ago, ruling out the 'late date proposal'¹³. Samples SR60C1 to SR62C, taken from still active flowstone covering ancient plaster, are all $< 2,500$ yr old, in agreement with the proposed age of the tunnel.

The historic-period dates of all Siloam Tunnel speleothems are compatible with our field observation that the Siloam Tunnel is human-made and did not follow a pre-existing karstic conduit^{17,24}. For comparison, we collected one flowstone sample (SR65) in the proximal Warren Shaft, which is a natural karstic shaft^{25,26}. Not surprisingly, this sample is older than the dating method limit (400 kyr), thus predating human construction of the waterworks by a considerable time span.

We have shown above that the ancient plaster was applied immediately after the Siloam Tunnel's construction. The material for plaster manufacture, including the dated organic matter, must have been readily available for the construction team. The piece of wood from 822–796 BC (SR53) retrieved from the ancient plaster had been originally taken from a tree that grew for some time before it was cut. Therefore it determines the maximum possible age of the Siloam Tunnel. Trees which reach a biological age of several hundreds years are rare in this region. Consequently, the Siloam Tunnel was probably constructed not later than ~ 100 yr after the 822–796 BC range of SR53, yielding a probable age for the Siloam Tunnel of about 700 BC or slightly earlier.

The age of a short-lived plant should be closer to the true archaeological date of construction. Indeed, the conventional ^{14}C age of the short-lived plant SR61 is $2,505 \pm 35$ yr, 115 yr younger than the age of SR53 ($2,620 \pm 35$ yr). The calendar age of SR61 is, with 95% probability, between 800 and 510 BC (the ^{14}C calibration curve is relatively flat at about 2,500 yr BP (ref. 21)). Within this range, the true age is most probably about 700 BC, as indicated by SR53 (see above).

On the basis of the ^{14}C dates it can be concluded that the Siloam Tunnel was most probably constructed in about 700 BC. This agrees well with the *terminus post quem* thorium age of the speleothems

deposited after the Siloam Tunnel's construction (see above).

The independent dating of both speleothem and plaster show that the Siloam Tunnel is hundreds of years older than the second century BC¹³. The radiometric Siloam Tunnel age of about 700 BC agrees well with the palaeographic age suggested for the Siloam Inscription^{14,15}. Our dating agrees well also with the date commonly assigned to King Hezekiah, whom the biblical text describes as having constructed the Siloam Tunnel. The three independent lines of evidence—radiometric dating, palaeography and the historical record—all converge on about 700 BC, rendering the Siloam Tunnel the best-dated Iron-Age biblical structure thus far known. □

Received 27 April; accepted 30 June 2003; doi:10.1038/nature01875.

- Finkelstein, I. & Silberman, N. *The Bible Unearthed: Archaeology's New Vision of Ancient Israel and the Origin of Sacred Texts* (Simon & Schuster, New York, 2001).
- Montgomery, J. A. *A Critical and Exegetical Commentary on The Book of Kings* (Scribner, New York, 1951).
- Vincent, L. H. *Underground Jerusalem: Discoveries on the Hill of Ophel (1909–11)* (Horace Cox, London, 1911).
- Amiran, R. in *Jerusalem Revealed* (ed. Yadin, Y.) 75–78 (Israel Exploration Society, Jerusalem, 1976).
- Ussishkin, D. The original length of the Siloam Tunnel in Jerusalem. *Levant* **8**, 82–95 (1976).
- Shaheen, N. The sinuous shape of Hezekiah's Tunnel. *Palestine Exploration Fund Quarterly Statement*, January–June, 103–108 (1979).
- Gill, D. Subterranean waterworks of Biblical Jerusalem: Adaptation of a karst system. *Science* **254**, 1467–1471 (1991).
- Abells, Z. & Arbib, A. *The City of David Water Systems* (Hadassah College, Jerusalem, 1995).
- Mazar, A. Jerusalem's water supply in the First Temple Period. in *The History of Jerusalem* (eds Ahituv, S. & Mazar, A.) 195–232 (Yad Izhak Ben-Zvi, Jerusalem, 2000).
- Reich, R. & Shukron, E. Reconsidering the karstic theory as an explanation to the cutting of Hezekiah's Tunnel in Jerusalem. *Bull. Am. Sch. Orient. Res.* **325**, 75–80 (2002).
- Issar, A. S. The evolution of the ancient water supply system in the region of Jerusalem. *Israel Explor. J.* **26**, 130–136 (1976).
- Tsuk, T. *Ancient Water Systems in Settlements in Eretz-Israel* (Tel Aviv Univ., 2000).
- Rogerson, J. & Davies, P. R. Was the Siloam Tunnel built by Hezekiah? *Bibl. Archaeologist* **59**, 138–149 (1996).
- Sasson, V. The Siloam Tunnel inscription. *Palestine Exploration Fund Quarterly Statement* **114**, 111–117 (1982).
- Della-Vida, G. L. in *In Memoriam Paul Kahle (BZAW103)* (eds Black, M. & Fohrer, G.) 162–166 (Topelmann, Berlin, 1968).
- Hackett, J. A. et al. Defusing pseudo-scholarship: The Siloam inscription ain't Hasmonean. *Bibl. Archaeol. Rev.*, March–April, 41–50 (1997).
- Shimron, A., Frumkin, A., Rosenbaum, J., Deutch, Y. & Dvoracek, M. The significance of geological structures, cave sediments and hydraulic lime in the interpretation of the subterranean waterworks beneath ancient Jerusalem. *Geol. Surv. Israel Curr. Res.* **12**, 247–256 (2000).
- Amit, D., Patrich, J. & Hirschfeld, Y. The aqueducts of Israel. *J. Roman Archaeol. Suppl. Ser. no. 46*, (Journal of Roman Archaeology, Portsmouth, Rhode Island, 2002).
- Hedges, R. E. M., Law, I. A., Bronk Ramsey, C. & Housley, R. A. The Oxford accelerator mass spectrometry facility: Technical developments in routine dating. *Radiocarbon* **31**, 99–113 (1989).
- Bronk Ramsey, C. Radiocarbon calibration and analysis of stratigraphy: The OxCal program. *Radiocarbon* **37**, 425–430 (1995).
- Stuiver, M. et al. INTCAL98 radiocarbon age calibration, 24,000–0 cal BP. *Radiocarbon* **40**, 1041–1083 (1998).
- Kaufman, A. et al. U–Th isotope systematics from the Soreq Cave, Israel and climatic correlations. *Earth Planet. Sci. Lett.* **156**, 141–155 (1998).
- Frumkin, A., Ford, D. C. & Schwarcz, H. P. Continental oxygen isotopic record of the last 170,000 years in Jerusalem. *Quat. Res.* **51**, 317–327 (1999).
- Frumkin, A. & Shimron, A. in *11th International Conference on Water in Antiquity* (eds Fahlbusch, H., Meshel, Z., Peleg, Y., Porath, Y. & Tsuk, T.) 4, (Tel Aviv University, 2001).
- Frumkin, A. *9th Congreso Internacional de Espeleología* 264–267 (Union Internacional de Espeleología, Barcelona, 1986).
- Gill, D. in *Excavations at the city of David 1978–1985 (Qedem 35)* (eds Ariel, D. T. & de Groot, A.) 1–28 (Institute of Archaeology, Hebrew University, Jerusalem, 1996).

Acknowledgements We thank U. Piner (Israel Cave Research Center) for assistance during the initial survey; S. Ashkenazi and Y. Peled (Geological Survey of Israel) for performing the drilling programme; the Research Laboratory for Archaeology and History of Art, Oxford University, for radiocarbon dating; the Open University, Milton Keynes, for $^{230}\text{Th}/^{234}\text{U}$ dating; C. Doherty (plaster and mortar) and P. Pettitt (radiocarbon) and Y. Elitzur for discussions, and K. Allred for surveying a new map of the Siloam Tunnel. The National Center for Cooperation Between Science and Archaeology provided financial assistance. Israel Antiquities Authority and Nature and Parks Authority permitted the study.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.F. (msamos@mscc.huji.ac.il).

Extreme reversed sexual size dimorphism in the extinct New Zealand moa *Dinornis*

Michael Bunce¹, Trevor H. Worthy², Tom Ford¹, Will Hoppitt¹, Eske Willerslev³, Alexei Drummond^{1,4} & Alan Cooper¹

¹Henry Wellcome Ancient Biomolecules Centre, Department of Zoology, University of Oxford, Oxford OX1 3PS, UK

²Palaeofaunal Surveys, 2A Willow Park Drive, RD11, Masterton, New Zealand

³Department of Evolutionary Biology, Zoological Institute, University of Copenhagen, 2100 Copenhagen, Denmark

⁴Department of Statistics, University of Oxford, Oxford OX1 3TG, UK

The ratite moa (Aves; Dinornithiformes) were massive graviportal browsers weighing up to 250 kg (ref. 1) that dominated the New Zealand biota until their extinction approximately 500 yr ago. Despite an extensive Quaternary fossil record, moa taxonomy remains problematic^{1–4} and currently 11 species are recognized. Three *Dinornis* species were found throughout New Zealand and differed markedly in size (1–2 m height at back) and mass (from ~34 to 242 kg)¹. Surprisingly, ancient mitochondrial DNA sequences show that the three species were genetically indistinguishable within each island, but formed separate North and South Island clades. Here we show, using the first sex-linked nuclear sequences from an extinct species, that on each island the three morphological forms actually represent just one species, whose size varied markedly according to sex and habitat. The largest females in this example of extreme reversed sexual size dimorphism were about 280% the weight and 150% the height of the largest males, which is unprecedented among birds and terrestrial mammals. The combination of molecular and palaeontological data highlights the difficulties of analysing extinct groups, even those with detailed fossil records.

In 1839 Owen predicted the existence of a large flightless ratite bird in New Zealand based on a fragmentary femur⁵ (Fig. 1a). The isolation and small size of New Zealand made the claim remarkable, although the unique, terrestrial mammal-free endemic fauna was dominated by a radiation of birds (an estimated 245 species¹).

Moa were adapted to a terrestrial, principally forest-dwelling habitat¹, and were highly variable morphologically resulting in the descriptions of over 64 species and 20 genera^{1,4}. Current schemes^{1,4} divide moa into two families (Emeidae, 8 species; Dinornithidae, 3 species). The short robust emeids differ from the tall and gracile dinornithids in many ways, notably in skull shape and vertebral structure, where the latter have three extra cervicals. The three *Dinornis* species (*Dinornis giganteus* Owen, 1844, *D. novaezealandiae* Owen, 1843 and *D. struthoides* Owen, 1844) show few cladistic differences, and throughout their history have been separated largely on the basis of limb bone size^{1–3,6,7} (Fig. 1a, b; see also Supplementary Information). An extensive Quaternary record indicates that *Dinornis* species had overlapping geographical ranges within New Zealand's North and South islands¹ (Fig. 1c).

To investigate the evolutionary history of *Dinornis*, polymerase chain reaction (PCR) primers were designed using complete mitochondrial genome sequences⁸. Rigorous ancient DNA techniques⁹ were used to sequence 525 base pairs (bp) of the control region from 32 *Dinornis* specimens, and 1,435 bp of protein-coding/transfer RNA sequences from seven of these (see Methods). Remarkably, despite the enormous morphological size variation there were no consistent genetic differences between the three *Dinornis* species. Furthermore, a strong phylogenetic signal separated all *Dinornis* specimens, irrespective of their size, into simple North Island or South Island clades (Fig. 2). This discrepancy between genetic

homogeneity and morphological variation could perhaps be explained by rapid morphological evolution, or differential growth patterns. Rapid dwarfing has been observed in other New Zealand avian groups over the past 10,000 years (10 kyr)¹, but most *Dinornis* sequences in Fig. 2 were from contemporaneous specimens (1–4 kyr), and all specimens were morphologically adult or fully grown (Supplementary Information). Differential growth by geographic area might have contributed some variation but the small *D. struthoides* co-occurred spatially and temporally with larger forms, *D. giganteus* and *D. novaezealandiae* (Fig. 1c).

In the absence of alternative explanations, the possibility that sexual dimorphism might be contributing to the large size variation in *Dinornis* was examined. In ratites, unlike other birds, genetic sexing is complicated because most of the female-specific W chromosome is genetically similar to the Z¹⁰. Recently, the nuclear-encoded W-specific KW1 locus has been used to sex modern ratite birds through a truncated W-specific copy with distinct sequence differences from multiple Z or autosomal copies¹¹. However, most ancient DNA studies have been restricted to high-copy number mitochondrial DNA sequences, and the amplification of single-copy nuclear loci is only expected in a subset of well-preserved ancient specimens¹². To maximize the number of moa specimens with sufficient nuclear DNA preservation for sex determination, it was necessary to design PCR primer pairs for a short

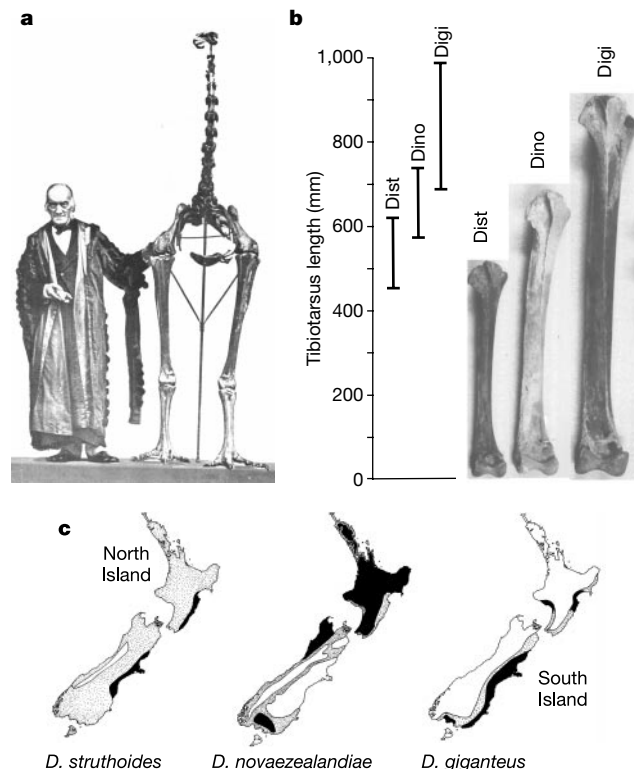


Figure 1 Size and distribution of the three currently accepted *Dinornis* species. **a**, Richard Owen holding the fossil femur fragment used to predict the existence of the moa in New Zealand. The skeleton (*D. novaezealandiae*) illustrates that *Dinornis* is the tallest bird known (ref. 24, pl. 47). **b**, Size comparison of the tibiotarsus of the three *Dinornis* species. The partly overlapping size ranges¹ illustrate the extent of variation in tibiotarsus length within each of the *Dinornis* species (Supplementary Information). Digi, Dino and Dist refer to *D. giganteus*, *D. novaezealandiae* and *D. struthoides*, respectively. **c**, Distribution of *Dinornis* Holocene palaeontological records within the North Island and South Island of New Zealand. Shading indicates prominence of species in fauna: black, prominent; stippled, regularly present; unshaded, rarely, if ever present.

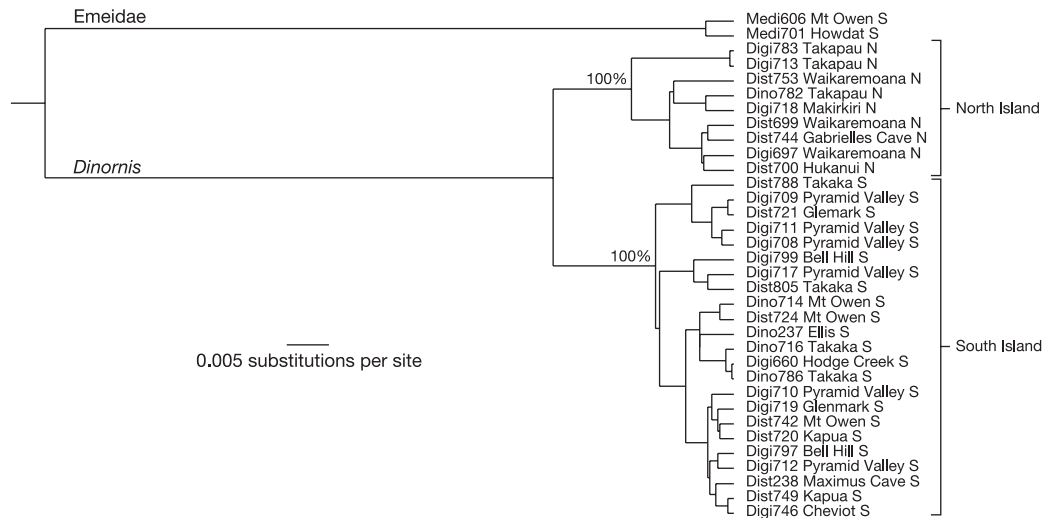


Figure 2 The maximum a posteriori tree of *Dinornis* mitochondrial DNA sequences generated from the posterior distribution using Metropolis–Hastings MCMC and control region sequences (525 bp) of 32 specimens, with protein-coding sequences (1,435 bp) from 7 specimens (see Methods). Two *Megalapteryx* moa sequences were included as out-groups. Morphological species identifications for *D. giganteus* (Digi),

D. novaezealandiae (Dino) and *D. struthoides* (Dist) are given, along with DNA extract numbers, collection sites and island of origin. All *Dinornis* specimens (irrespective of species) fall into either the North (N) or South Island (S) clade. The split between the islands has a posterior probability of 100%, and is estimated to be mid-Pleistocene in origin (see Supplementary Information).

W-specific (female) product. KW1 sequences of modern rhea, ostrich and kiwi samples were generated and revealed several differences from those previously published¹¹, as well as several extra Z or autosomal copies (data not shown). The primer pair 112f and 267r (Supplementary Information) consistently produced a 180-bp product from female, but not male, samples of rhea, ostrich and kiwi (Fig. 3a; see also Methods). A similar-sized amplification product was also generated in several well-preserved *D. giganteus* and *D. novaezealandiae* specimens but not in any *D. struthoides*

samples. The sequence of the moa 112f–267r product was similar to other ratite W-specific KW1 sequences (Supplementary Information), and was used to design internal primers (185f and 260r) for a short (85 bp) moa W-specific product suitable for amplification from degraded ancient templates (Fig. 3b).

The 185f and 260r KW1 primers were used to screen 34 *Dinornis* specimens, along with a similar-sized control amplification (85 bp) of the nuclear-encoded alcohol dehydrogenase (*ADH*) gene (Fig. 3b). Eight specimens (24%) yielded neither the ADH-positive

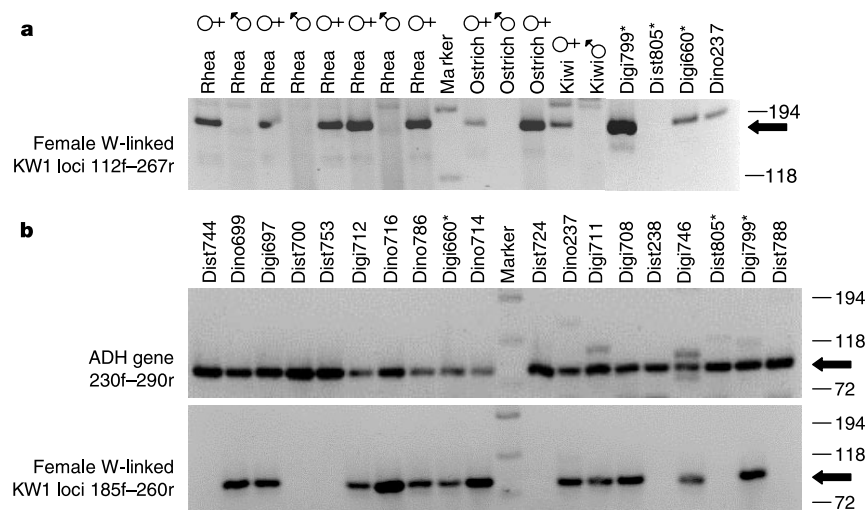


Figure 3 Molecular sex determination of *Dinornis*. **a**, A 180-bp W-linked product was produced from the modern female rhea, ostrich and kiwi samples using the 112f and 267r primer combination. All male samples failed to amplify a product of the correct size (indicated by arrows) or sequence. Similar bands and sequences are shown from two well-preserved *D. giganteus* (Digi799, Digi660) and *D. novaezealandiae* (Dino237) specimens, but not in a well-preserved *D. struthoides* (Dist805) specimen. **b**, Sub-fossil

Dinornis samples screened for short (85-bp) W-specific and ADH products. ADH PCR products were obtained from 26 specimens (19 shown, upper panel), of which 15 of the 16 *D. giganteus* and *D. novaezealandiae* samples also tested positive for the presence of the 85-bp W-specific amplification product (12 shown, lower panel). Of the 10 *D. struthoides* samples positive for ADH, none gave a W-linked product. The three samples marked with an asterisk were independently replicated in Copenhagen.

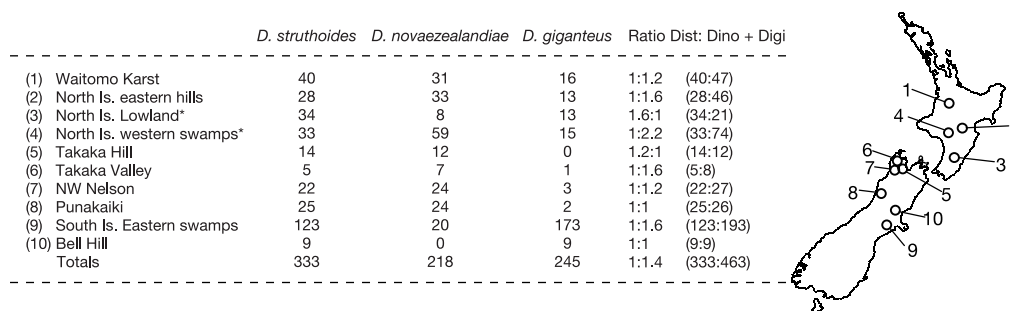


Figure 4 Occurrence of *Dinornis* skeletal remains from swamp and cave deposits across New Zealand. Palaeontological records show that the ratio of *D. struthoides*:*D. novaezealandiae* plus *D. giganteus* (male:female) is 1:1.4 across ten broad geographical areas for which detailed data exists, which is within the range of modern ratites (1:1.1–

1.44). Only two swamp-dominated sites (asterisks) differ significantly from the 1:1.4 ratio (χ^2 test; $P > 0.2$), and these may reflect taphonomic biases such as weight-based entrapment or seasonality.

control nor female-specific KW1 products, presumably owing to a lack of amplifiable nuclear DNA. The ADH locus was amplified from the remaining 26 specimens (10 *D. giganteus*, 6 *D. novaezealandiae* and 10 *D. struthoides*), of which 15 of the 16 *D. giganteus* and *D. novaezealandiae* samples were also positive for the female-specific product. In direct contrast, none of the 10 *D. struthoides* samples positive for ADH gave W-specific products (Fig. 3b). Determination of relative numbers of nuclear templates in six extracts by quantitative real-time PCR confirmed *D. struthoides* specimens were among some of the best preserved (Supplementary Information). Consequently, the failure of all the *D. struthoides* specimens to produce female-specific KW1 products seems unlikely to reflect PCR drop-out. The one *D. giganteus* extract (Digi783) that yielded an ADH, but not a KW1 product, showed high levels of PCR inhibition in the quantitative assays, explaining the inconsistent amplification results. Independent analysis of three specimens in Copenhagen replicated the KW1 and ADH results (Fig. 3b).

The mitochondrial data suggest that a single species of *Dinornis* existed on each island, whereas the nuclear sequences indicate that the larger specimens (previously identified as *D. giganteus* and *D. novaezealandiae*) were females and the smaller specimens (*D. struthoides*) were male. A re-examination of the palaeontological data provides further support for such extreme reversed sexual dimorphism (RSD). The distribution of *Dinornis* specimens in well-characterized swamp and cave deposits across New Zealand (Fig. 4) reveals that when *D. giganteus* and *D. novaezealandiae* are considered together (as females), the ratio of putative male (*D. struthoides*) to female (*D. novaezealandiae* and *D. giganteus*) averages 1:1.4 (Fig. 4), within the range of modern ratites (1:1–1.44)^{13,14}. Only two swamp deposits depart markedly from this ratio and these may be the consequence of taphonomic biases such as mass-dependent entrapment and/or seasonal variation in swamp size. The size of female *Dinornis* (*D. giganteus* and *D. novaezealandiae*) appears to have ranged from around 1.2–1.9 m in height (at back) and 76–242 kg in mass (Supplementary Information). Both size and mass varied according to habitat, with females averaging 102 kg in northwest South Island and 167 kg in eastern South Island. Males (*D. struthoides*) were significantly smaller, ranging from 0.9–1.2 m and 34–85 kg. Smaller size in both sexes appears to correlate with higher rainfall, more homogeneous closed canopy vegetation, higher altitude, and may relate to available nutritional content of vegetation.

The combined molecular and palaeontological data strongly support RSD in *Dinornis* on a scale unprecedented among birds or terrestrial mammals, with the largest females about 150% the height and 280% the weight of the largest males (Supplementary Information). Other forest-dwelling ratites, such as the kiwi (*Apteryx*) and southern cassowary (*Casuaris casuaris*) exhibit RSD, with females typically 120% and up to 180% the weight of

males, respectively^{13,15}, whereas two species of emeid moa have also been suggested to show sexual dimorphism¹⁶. Interestingly, RSD is much less evident or non-existent in plain-dwelling ratites. In general, avian RSD is associated with increased investment in egg production¹⁷ and kiwi are a good example. Ecological factors, such as niche partitioning, have also been suggested to exist in *Dinornis* taxa^{18,19}. Overall, both the RSD and absolute size variation observed in *Dinornis* seem to represent extreme examples of trends seen elsewhere.

Molecular rate analysis indicates that the North Island and South Island *Dinornis* mitochondrial lineages separated around the mid-Pleistocene epoch, consistent with geological activity in Cook Strait²⁰, the marine barrier between the islands (Fig. 2; see also Methods). The reciprocal monophyly of the mitochondrial DNA lineages from the two island populations makes it unlikely that substantial gene flow has occurred since then, or that the genetic separation has been caused by geographical lineage sorting²¹. This implies that the existence of RSD on each island represents either convergence or an ancestrally inherited character. Either way, the resolution of all *Dinornis* forms into at most two allopatric taxa explains the lack of cladistic morphological characters separating the previously identified taxonomic groups. The following taxonomy is now advocated: family, Dinornithidae (Bonaparte); genus, *Dinornis* (Owen); species, *D. novaezealandiae* Owen, 1843 (North Island *Dinornis*) and *D. robustus* Owen, 1846 (South Island *Dinornis*). The species *D. giganteus* and *D. struthoides* are placed in the synonymy of *D. novaezealandiae*.

The first molecular sexing of an extinct species has revealed an example of extreme RSD which was previously unsuspected despite an extensive Quaternary record. This combination of molecular and palaeontological data clearly illustrates the potential difficulties of characterizing extinct taxa with less detailed fossil records. □

Methods

Sample information

Sections of cortical bone (about 0.1 g) were collected from moa leg bones and prepared in a physically isolated, ancient DNA laboratory at the Oxford University Museum of Natural History. Museum numbers, provenance, site/bone dates and bone morphology appear in Supplementary Information. Length data for *Dinornis* femora, tibiotarsi and tarsometatarsi were obtained as previously described⁷ for all sites with well-characterized material¹⁷ (Supplementary Table 1).

DNA extraction, amplification and sequencing

DNA was extracted and amplified as previously described^{8,21} using appropriate ancient-DNA techniques⁹ in a dedicated laboratory. Multiple negative extraction and amplification controls were included, and selected amplification products were cloned using TOPO TA cloning kits (Invitrogen) to detect contamination and DNA damage. Modern samples and ancient samples subsequent to PCR amplification were analysed in the Zoology Department. Three moa bone samples (Digi799, Digi660 and Dist805) were sent to the University of Copenhagen for independent replication. All PCR reactions were conducted as described²¹ using Hi-fidelity Platinum Taq (Invitrogen) to minimize polymerase error. Primers were designed from mitochondrial⁸, ADH²² and KW1¹²

sequences and are given in Supplementary Information. Thermal cycling conditions were 40 cycles of 95 °C/55–60 °C/68 °C (30–45 s each) for mitochondrial amplifications and 40–50 cycles of 95 °C/52–55 °C/68 °C (45 s each) for KW1 and ADH amplifications respectively. PCR re-amplifications were only occasionally required, and used the same primers. Sequences were determined using ABI Big Dye (v.3) on an ABI 310 or 3700 (Applied Biosystems), according to manufacturers instructions.

Phylogenetic methods

Metropolis–Hastings Markov chain Monte Carlo (MCMC) integration was used to jointly estimate the phylogenetic tree and parameters of the substitution model under the assumption of a molecular clock²³. An HKY + G + I model of substitution was used to characterize the evolution of the concatenated control region and protein-coding sequences (525-bp control region for 32 *Dinornis* specimens, Medi606, Medi701, and 1435 bp of COI, II, tSer, tAsp, tLys, and ATP8 genes for Digi660, Dist700, Dist238, Dino699, Dino237, Digi697, Dist753, Medi606, Medi701). Analysis of the control region sequences alone produced almost identical estimates (Supplementary Information). The tree presented is the maximum a posteriori tree obtained from the MCMC analysis. The benefit of the MCMC approach lies in assessing the uncertainty in the tree through bayesian posterior probabilities of the relevant clades.

Received 6 March; accepted 23 June 2003; doi:10.1038/nature01871.

1. Worthy, T. H. & Holdaway, R. N. *The Lost World of the Moa* (Indiana Univ. Press, Bloomington, 2002).
2. Archey, G. The moa: A study of the Dinornithiformes. *Bull. Auckland Mus.* 1, 1–144 (1941).
3. Oliver, W. R. B. The moas of New Zealand and Australia. *Dominion Mus. Bull. Well. NZ*, 15, 1–204 (1949).
4. Cracraft, J. Covariation patterns in the post-cranial skeleton of the moas (Aves, Dinornithidae): A factor analytic study. *Paleobiology* 2, 166–173 (1976).
5. Owen, R. On the bone of an unknown struthious bird from New Zealand, meeting of November 12, 1839. *Proc. Zool. Soc. Lond.* 7, 169–171 (1840).
6. Hutton, F. W. The moas of New Zealand. *Trans. Proc. NZ Inst.* 24, 93–172 (1892).
7. Worthy, T. H. Reappraisal of *Dinornis* (Aves, Dinornithiformes) species—a morphometric analysis. *NZ J. Zool.* 21, 113–134 (1994).
8. Cooper, A. et al. Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. *Nature* 409, 704–707 (2001).
9. Cooper, A. & Poinar, H. N. Ancient DNA: Do it right or not at all. *Science* 289, 1139 (2000).
10. Shetty, S., Griffin, D. K. & Graves, J. A. M. Comparative painting reveals strong chromosome homology over 80 million years of bird evolution. *Chromosome Res.* 7, 289–295 (1999).
11. Huynen, L., Millar, C. D. & Lambert, D. M. A DNA test to sex ratite birds. *Mol. Ecol.* 11, 851–856 (2002).
12. Greenwood, A. D., Capelli, C., Possnert, G. & Paabo, S. Nuclear DNA sequences from late Pleistocene megafauna. *Mol. Biol. Evol.* 16, 1466–1473 (1999).
13. Davies, S. J. J. F. *Ratites and Tinamous* (Oxford Univ. Press, Oxford, 2002).
14. Bertram, B. C. R. *Breeding Systems and Strategies of Ostriches* 890–894 (Proc. XVII Int. Ornith. Congr., Berlin, 1979).
15. Marchant, M. & Higgins, P. J. *Handbook of Australian, New Zealand and Antarctic Birds* (Oxford Univ. Press, Melbourne, 1990).
16. Worthy, T. H. Sexual dimorphism and temporal variation in the North Island moa species *Euryapteryx curtus* (Owen) and *Pachyornis mappini* (Archey). *Natl Mus. NZ Rec.* 3, 59–70 (1987).
17. Andersson, M. *Sexual Selection* (Princeton Univ. Press, 1994).
18. Atkinson, I. A. E. & Greenwood, R. M. Relationships between moas and plants. *NZ J. Ecol.* 12, 67–96 (1989).
19. Shine, R. Ecological causes for the evolution of sexual dimorphism—a review of the evidence. *Q. Rev. Biol.* 64, 419–461 (1989).
20. Lewis, K. B., Carter, L. & Davey, F. J. The opening of Cook Strait—interglacial tidal scour and aligning basins at a subduction to transform plate edge. *Mar. Geol.* 116, 293–312 (1994).
21. Barnes, I., Mathews, P., Shapiro, B., Jensen, D. & Cooper, A. Dynamics of Pleistocene population extinctions in Beringian brown bears. *Science* 295, 2267–2270 (2002).
22. Hjeltnist, L. et al. Alcohol-dehydrogenase of class-I—kiwi liver-enzyme, parallel evolution in separate vertebrate lines, and correlation with 12s ribosomal-RNA patterns. *FEBS Lett.* 367, 306–310 (1995).
23. Drummond, A. J., Nicholls, G. K., Rodrigo, A. G. & Solomon, W. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* 161, 1307–1320 (2002).
24. Owen, R. *Memoirs on the Extinct Wingless Birds of New Zealand* (London, 1879).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank the Museum of New Zealand Te Papa Tongarewa (A. J. D. Tennyson, J. A. Bartle), Canterbury Museum (P. Scofield), Auckland Museum (B. Gill), Otago Museum (S. Michelsen-Heath), B. Reeve, Bell Hill Vineyard (M. Giesen, S. Veldhuizen), Pampas Poultry (H. Macfie), B. Shapiro, T. Gilbert, R. Holdaway, J. Binladen and T. Brand for samples and assistance with analysis. The Oxford University Museum of Natural History provided laboratory space. We thank NERC (A.C. and M.B.), Wellcome and Leverhulme Trusts (A.C.), BBSRC (T.F., W.H., A.D., A.C.), EPSRC (A.D.), Villum Kann Rasmussen Fonden, Denmark (E.W.) and the New Zealand ForST (T.H.W.) for financial support.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.C. (alan.cooper@zoo.ox.ac.uk). Sequences have been deposited in GenBank under accession numbers AY326127–AY326193.

Nuclear DNA sequences detect species limits in ancient moa

L. Huynen¹, C. D. Millar², R. P. Scofield³ & D. M. Lambert¹

¹Allan Wilson Centre for Molecular Ecology and Evolution, Institute of Molecular BioSciences, Massey University, Private Bag 102 904, and ²Allan Wilson Centre for Molecular Ecology and Evolution, School of Biological Sciences, University of Auckland, Private Bag 92019, Auckland, New Zealand

³Canterbury Museum, Rolleston Avenue, Christchurch 8001, New Zealand

Ancient DNA studies have typically used multi-copy mitochondrial DNA sequences^{1,2}. This is largely because single-locus nuclear genes have been difficult to recover from sub-fossil material³, restricting the scope of ancient DNA research. Here, we have isolated single-locus nuclear DNA markers to assign the sex of 115 extinct moa and, in combination with a mitochondrial DNA phylogeny, tested competing hypotheses about the specific status of moa taxa. Moa were large ratite birds that showed extreme size variation both within and among species⁴. For some taxa, this large variation was hypothesized to represent sexual dimorphism, while for others it was argued to reflect the existence of different species⁵. Our results show that moa were characterized by extreme reverse sexual dimorphism and as a result we have been able to clarify the number of moa species. For example, we show that the three recognized ‘species’ of *Dinornis* comprised only two monophyletic groups and that two of these ‘species’ comprised individuals of one sex only. This study also illustrates that single-locus nuclear DNA sequences can be consistently recovered from ancient material.

There are typically hundreds of copies of the mitochondrial genome in most cells, while the nuclear genome is usually present in only two copies⁶. Owing to this difference in copy number, mitochondrial genes are more likely to survive in ancient material⁷. Consequently, there has been an almost complete reliance on mitochondrial DNA sequences in studies of extinct organisms^{8–10}. If limited-copy nuclear genes could be consistently recovered, they would provide the opportunity to address a series of previously intractable questions in ancient ecology and evolution. For example, species limits in many extinct vertebrates have been determined by reference to variation in the morphology of skeletal remains^{11,12}. However, this approach is problematic in the case of the moa of New Zealand¹³, as species delimitation has been confounded by high levels of morphological variation within and among taxa.

Early skeletal analyses, originating with descriptions by Polack in 1838¹⁴ and Owen in 1839¹⁵, suggested that there were very many species of moa, with at least 64 species names being used, together with 20 generic names¹⁶. Over the last 25 yr the number of recognized species has reduced substantially from 38 to 11 (refs 5, 16). The smaller number is primarily based on the hypothesis that some previously recognized ‘species’ that differed in size only, actually represented conspecific males and females, with females being the larger sex. For example, in the currently recognized species *Pachyornis mappini*, *Emeus crassus*, and *Euryapteryx curtus* a bi-modal size distribution is evident and has been interpreted as intra-specific sexual dimorphism⁵. Other taxa were suggested to represent separate species on the basis of what were regarded as acceptable limits of intra-specific size variation, using living ratite species as a guide. In the heaviest of moa, *Dinornis*, size variation was shown to be exceptionally large and so three morphologically similar, but differently sized species have been designated: *D. giganteus*, *D. novaezealandiae*, and *D. struthoides*. As a result of these osteological analyses, it was concluded that there were, in fact, only 11 species of moa^{5,16}.

We tested the hypothesis that sexual dimorphism was prevalent in

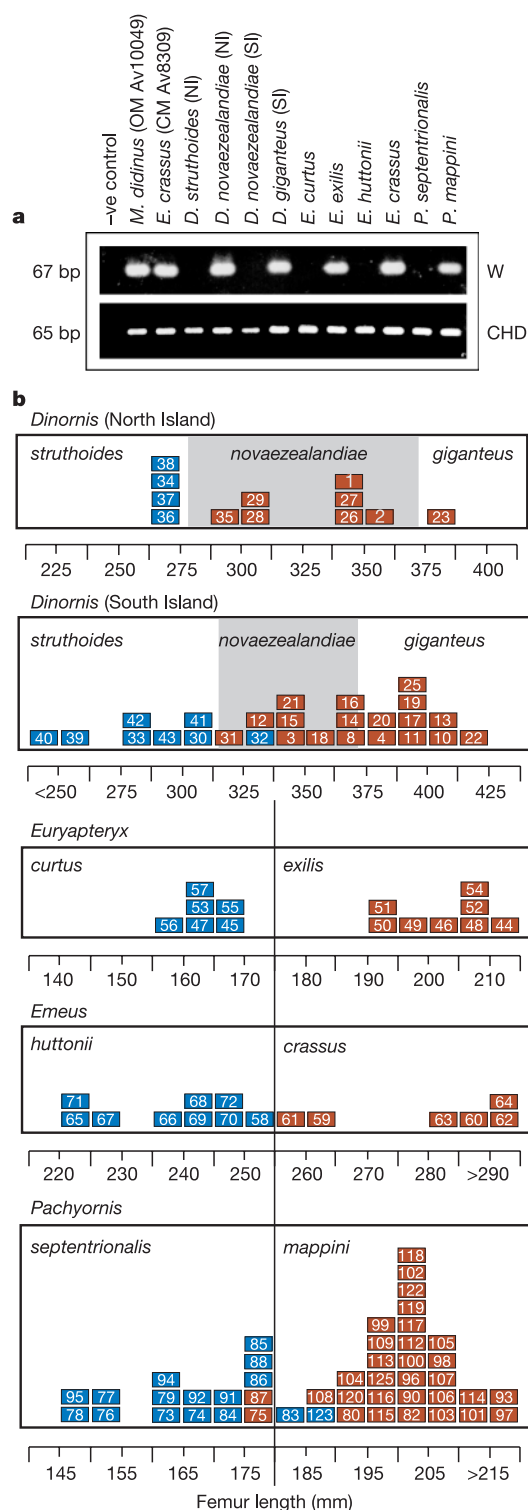


Figure 1 Molecular sexing of moa. **a**, DNA samples were sexed by amplification of single locus W chromosome sequences using primers moa1 and moa2. Females are identified by the presence of a 67-bp PCR product. Amplification of 65-bp sequence of the nuclear CHD gene, using primers pM0 and pM1, served as an amplification control. DNA extracted from the known sex *E. crassus* (CM Av8309) and *M. didinus* (OM Av10049) served as positive controls. NI and SI refer to the North Island and South Island of New Zealand. **b**, Moa sub-fossil bones are classified according to size, as detailed in refs 5, 24. Numbers in boxes correspond to sample numbers shown in Fig. 2. Red boxes indicate females and blue boxes indicate males, based on DNA analyses.

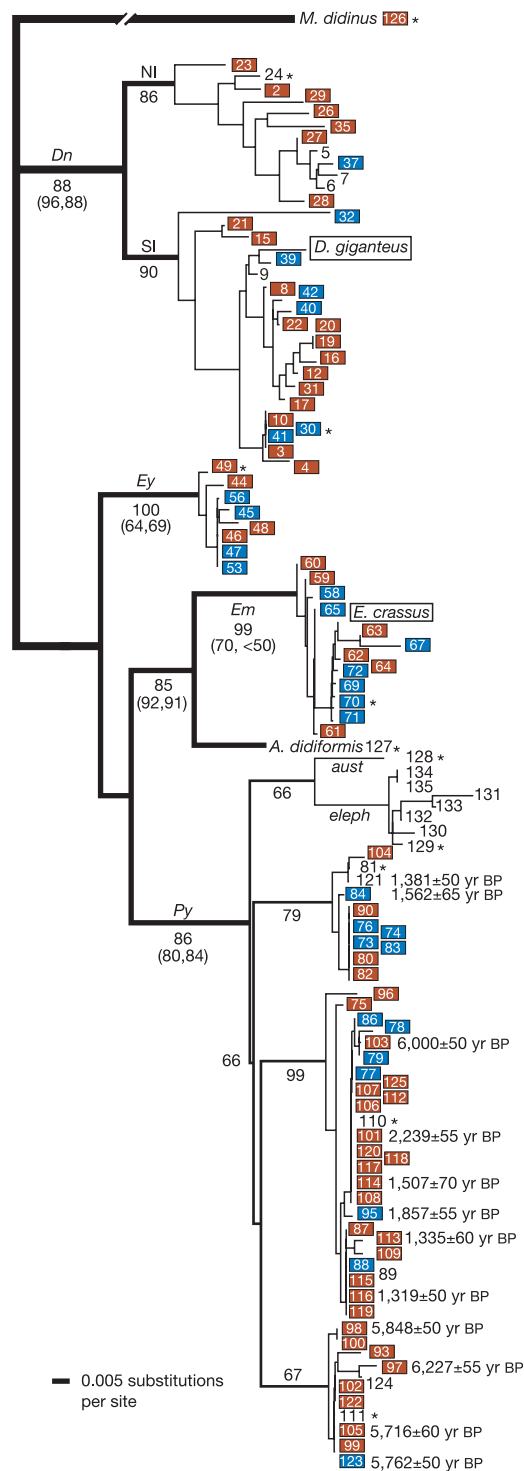


Figure 2 Phylogenetic relationships among currently recognized moa taxa. Each individual is represented by a number (with corresponding details in the Supplementary Information). A neighbour-joining (distance) tree was constructed from an ~240-bp sequence of 5' mitochondrial HVRI region (GenBank accession numbers: AY299860–AY299971). The tree is rooted with *M. didinus*. Previously published moa sequences (in boxes) for *E. crassus* (AY016015)⁸ and *D. giganteus* (AY016013)⁸ are also included. Bootstrap values were calculated from 1,000 re-samplings, using neighbour-joining analysis and Kimura's two-parameter model of substitution. Maximum parsimony and maximum likelihood bootstrap values, using representatives of each clade (asterisks) are shown in brackets. The red and blue boxes indicate females and males respectively. Species (abbreviations): *Emeus* (Em), *Euryapteryx* (Ey), *Pachyornis* (Py) and *Dinornis* (Dn).

moa, by amplifying single-copy sex-linked DNA sequences. All birds have a ZW/ZZ chromosomal sex-determining mechanism, with females (ZW) being the heterogametic sex^{17,18}. We have previously isolated W chromosome DNA sequences from extant ratites and have shown that polymerase chain reaction (PCR) primers designed to these sequences allow the assignment of sex in all living ratites¹⁹. For the present ancient DNA study, extensive sequencing of the W chromosome of a number of moa species enabled the design of PCR primers that can be used to assign the sex of moa. These primers correctly sexed the only two known-female moa specimens, one of *E. crassus* (sample 60), and a sub-fossil bone of *Megalapteryx didinus* (sample 126; S. Heath, personal communication; Fig. 1a).

The single-locus nuclear gene for CHD (chromo-helicase DNA-binding protein) was used as a positive control in all amplifications^{20,21} (Fig. 1a). All DNA extractions were carried out in a physically separate laboratory dedicated to the study of ancient DNA at Massey University, Palmerston North, New Zealand. Sequences were obtained from multiple amplifications of individual samples. In addition, PCR products from seven sub-fossil bone samples were cloned and the sequences of at least eight clones determined for each sample. In each case the consensus sequence of the clones was identical to that of the PCR products (Supplementary Information). DNA extractions, amplifications and sequences were duplicated for seven femora at an ancient-DNA facility at the University of Auckland, New Zealand. All sequences were identical to those previously obtained. For sexing, DNA samples from 42 sub-fossil femora were amplified at least six times to ensure consistency^{22,23}. A total of 115 birds were sexed, comprising 72 females and 43 males. This represents a significant excess of females ($\chi^2 = 7.31$, $P > 0.05$), and suggests that our test does not mis-score females owing to non-amplification of W chromosome sequences. Details of the number of separate W chromosome-specific amplifications are given in the Supplementary Information. By amplification of a small (65 base pairs, bp) W chromosome fragment, we were able to sex more than 91% of all samples used for mitochondrial sequence analysis. Bones from these sites ranged in radiocarbon age from 1,319–6,227 yr BP. We have been able to successfully recover single-locus nuclear DNA sequences of up to 200 bp from these bones (data not shown) and consequently, we suggest that nuclear DNA sequences should be recoverable from sub-fossils of a considerable age.

The following pairs of taxa, *E. curtus*/*E. exilis*, *E. huttonii*/*E. crassus*, and *P. septentrionalis*/*P. mappini*, had previously been reduced, in each case, to single species⁵. This was because the members of each pair were hypothesized by Cracraft⁵ to represent either males or females. In each case, we found a clear relationship

between the size of femora and the presence of W chromosome DNA sequences (Fig. 1b). That is, these species were sexually dimorphic, with females consistently being the larger sex, which suggests that the proposed reduction in the number of species is warranted. On the basis of bone morphology alone, distinct non-overlapping size ranges have been suggested for males and females of *Emeus*, *Euryapteryx*, and *Pachyornis*²⁴. By using a large sample size, we show that for *P. septentrionalis*/*P. mappini* an overlap in the size range of males and females is evident. This is likely to become more pronounced as sample size increases in future studies. Furthermore, a bias in the sex ratio of *P. septentrionalis*/*P. mappini*, with an excess of females (25 females and 4 males), was detected in samples from two North Island swamp sites: Makirikiri and Riverlands (Supplementary Information). This result supports suggestions¹⁶ that females were more likely to be trapped in swamp environments, perhaps as a result of their larger size and/or more adventurous foraging behaviour.

The phylogenetic relationships among moa taxa were determined by amplifying and sequencing approximately 240 or 440 bp of DNA from the 5' terminus of the mitochondrial control region (hyper-variable region I). These amplifications were performed on sub-fossil bones, using primers specifically designed for moa. Approximately 440 bp of mitochondrial sequence was obtained from two overlapping DNA fragments, each approximately 240 bp in length (see Methods). The relationships among moa sequences, in the form of a neighbour-joining tree, together with the sex of individuals and radiocarbon ages of bones are shown in Fig. 2. Maximum parsimony and maximum likelihood analyses of representative samples resulted in phylogenetic trees with virtually identical topology. *M. didinus* was used as the outgroup in all analyses^{16,25}.

The suggestion that *E. curtus*/*E. exilis*, *E. huttonii*/*E. crassus*, and *P. septentrionalis*/*P. mappini*, represent conspecific males and females is also supported by the molecular phylogeny (Fig. 2). Individuals from each pair of taxa are dispersed throughout monophyletic clades within the tree. In the case of *Pachyornis*, three clades are evident and individuals identified as *P. mappini* and *P. septentrionalis* are found distributed throughout each.

Unexpectedly, the three currently recognized 'species' of *Dinornis* are shown to comprise only two monophyletic clades; one from the North Island and one from the South Island (Fig. 2). We carried out a further analysis of this unusual pattern by sequencing an additional 203 bp of flanking control region DNA, with virtually identical results to those presented in Fig. 2 (Supplementary Fig. S2). Sequences from the three 'species' are found in both the North Island and South Island clades. That is, sequences from *D. giganteus*, *D. novaezealandiae*, and *D. struthoides* from each island form two monophyletic assemblages, rather than grouping with samples of

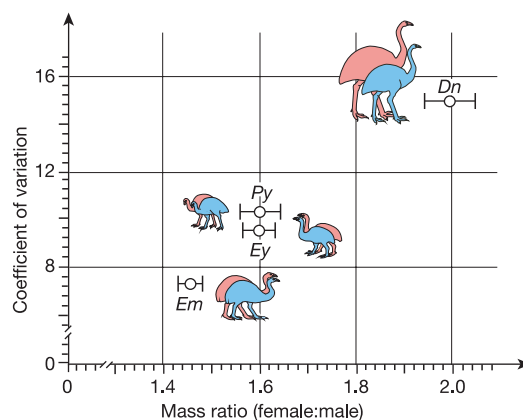


Figure 3 Comparison of coefficients of variation for femora lengths with female: male mass, in moa. The relative size of male (blue) and female (red) moa are given for four

species (abbreviations as in Fig. 2). Moa mass was calculated using femur length according to ref. 26.

the same taxonomically recognized species. Figure 1b shows the pattern of sex-specific DNA amplifications for *Dinornis* specimens in relation to femora lengths. Surprisingly, the single North Island femur of *D. giganteus* and the material from the seven North Island *D. novaezealandiae* examined were from females. Similarly, South Island *D. giganteus* and most *D. novaezealandiae* were females, and in both islands all specimens of *D. struthoides* examined were males. In combination with the phylogeny, these results show that only one *Dinornis* species was present in each island. The size range of specimens from the South Island is exceptionally large (coefficient of variation, 15.1), in comparison to the relatively low levels of variation present within other moa species (coefficients of variation 7–10.7; Fig. 3). The data in Fig. 3 show that this high level of variation in *Dinornis* is not the result of a large size range within males and females, but is due to a large size difference between the two sexes. In fact, estimates of average male and female mass, using femur length²⁶, suggests that *Dinornis* females were almost twice the mass of their male counterparts (Fig. 3).

The original designation of three *Dinornis* species was based on the assumption that the level of size variation detected in this genus is too large to represent a single species. Collectively, our data suggest that the three *Dinornis* 'species' are not biologically distinct, and that there were, at most, two geographically isolated species of *Dinornis* in New Zealand during the Holocene period. Furthermore, our results show that levels of sexual dimorphism can vary greatly, among even closely related species, and that such variation may confound estimates of species limits.

Thus our isolation of single locus nuclear DNA markers for moa sex, together with phylogenies of a rapidly evolving region of the mitochondrial genome²⁷, has enabled us to determine species limits in moa. More generally, this study illustrates the role that single-copy nuclear DNA sequences can play in testing previously intractable hypotheses about extinct biota. □

Methods

DNA extraction, PCR, sequencing

Moa samples were obtained from the following museums: Auckland, Waitomo Caves, Whanganui Regional, Te Manawa (Palmerston North), Museum of New Zealand Te Papa Tongarewa (Wellington), Canterbury, and Otago. Femora were drilled using an 8-mm drill bit. The drill bit was sterilized in 5% (v/v) NaHClO between uses and initial shavings obtained from the bone surface were discarded. DNA was extracted from approximately 0.1 g of bone shavings by rotation overnight at 55 °C in 3 ml of 0.5 M EDTA (pH 8.0) with approximately 2 mg of proteinase K. The DNA was purified by phenol/phenol:chloroform:isoamyl alcohol/butan-2-ol extraction²⁸, then concentrated and washed using Vivaspin concentrators (30 000 MWCO PES; Vivascience). All DNA amplifications were carried out in 10 µl reactions containing 50 mM Tris-HCl pH 8.8, 20 mM NH₄SO₄, 2.5 mM MgCl₂, 100 µM of each dNTP, 0.3 U AmpliTaq DNA polymerase (Perkin Elmer), 40 ng of each primer, and approximately 1–20 ng of DNA. Approximately 240 bp of DNA sequence from the 5' terminus of the mitochondrial control region was amplified using primers EC2 (5'-TAGACCTTACAGCACTACATAGGAA-3') and RmoaCRrev (5'-CATTTGGCTCATACCCATC-3'). A further 203 bp of flanking sequence was obtained using the primers moaCR8 (5'-GCATAGATTTATAGCTCGGACAT-3') and moaCR4 (5'-AATCTATTAATGACCTCGACTTAGGA-3'). Reactions were heated to 94 °C for 2 min, and then subjected to ten cycles of 94 °C for 20 s, 52 °C for 20 s and 72 °C for 30 s, and then 35 cycles of 94 °C for 20 s, 50 °C for 20 s, and 72 °C for 30 s. For weak amplifications, a second PCR was carried out by slicing the PCR product from Tris-acetate-buffered agarose (Boehringer-Mannheim), centrifuging the gel slice through a 200-µl barrier tip (Axxygen Scientific), and using 0.5 µl of the flow-through in a fresh amplification mix. This mix was then subjected to 25 cycles of 94 °C for 2 min, 52 °C for 20 s, and 72 °C for 20 s. Mitochondrial PCR products were visualized in 1% MS/1% LE agarose (in Tris-acetate buffer), then purified using a High Pure PCR Product Purification Kit (Roche) and sequenced using ABI fluorescent dye terminator cycle sequencing chemistry. Sequence reactions were analysed using an ABI377 automated sequencer.

Genetic sexing

Femora were sexed using the primers moa1 (5'-CCATGTTCACTGTTTCTTACTAAT-3') and moa2 (5'-ATGTTAAGCAATGCTCTATGACAGA-3'). Reactions were heated to 94 °C before the addition of the primers and then subjected to 45 cycles of 94 °C for 20 s, 58 °C for 20 s and 72 °C for 20 s. The presence of nuclear DNA was determined by amplification of 65 bp of the nuclear gene *CHD* using primers pM0 (5'-GATAGTGACTCCATCTCAGAA-3') and pM1 (5'-GGGAATAGTTCGTGGTCTTCC-3'). These primers were designed from moa sequences obtained using primers p2 and p3²⁹. Amplification of *CHD* sequences was performed as for moa1 and moa2, but with the annealing

temperature set to 60 °C. Nuclear DNA products were visualized in 2% MS/1% LE agarose in Tris-borate buffer.

Radiocarbon dating

Radiocarbon dates of moa bones used in this study were determined using accelerated mass spectrometry (AMS) to measure the abundance of the cosmogenic radioisotope ¹⁴C. Radiocarbon ages were supplied by the Institute of Geological and Nuclear Sciences, Lower Hutt, New Zealand.

Phylogenetic analysis

Mitochondrial sequences were initially aligned using Sequencher 4.1 and then realigned by eye. Phylogenetic analyses were carried out using the program PAUP* 4.0b10 (ref. 30).

Received 14 March; accepted 3 June 2003; doi:10.1038/nature01838.

- Higuchi, R. *et al.* DNA sequences from the quagga, an extinct member of the horse family. *Nature* **312**, 282–284 (1984).
- Herrmann, B. & Hummel, S. (eds) *Ancient DNA* (Springer, New York, 1993).
- Cooper, A. & Wayne, R. New uses for old DNA. *Curr. Opin. Biotechnol.* **9**, 49–53 (1998).
- Anderson, A. J. *Prodigious Birds: Moas and Moa-Hunting in Prehistoric New Zealand* (Cambridge Univ. Press, Cambridge, UK, 1989).
- Cracraft, J. The species of moa. *Smithson. Contrib. Paleobiol.* **27**, 189–205 (1976).
- Pikö, L. & Taylor, K. D. Amounts of mitochondrial DNA and abundance of some mitochondrial gene transcripts in early mouse embryos. *Dev. Biol.* **123**, 364–374 (1987).
- Hofreiter, M., Serre, D., Poinar, H. N., Kuch, M. & Pääbo, S. Ancient DNA. *Nature Rev. Genet.* **2**, 353–359 (2001).
- Cooper, A. *et al.* Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. *Nature* **409**, 704–707 (2001).
- Haddrath, O. & Baker, A. J. Complete mitochondrial DNA genome sequences of extinct birds: ratite phylogenetics and the vicariance biogeography hypothesis. *Proc. R. Soc. Lond. B* **268**, 939–945 (2001).
- Leonard, J. A. *et al.* Ancient DNA evidence for Old World origin of New World dogs. *Science* **298**, 1613–1616 (2002).
- Asfaw, B. *et al.* Remains of *Homo erectus* from Bouri, Middle Awash, Ethiopia. *Nature* **416**, 317–320 (2002).
- Weins, J. & Penkrot, T. Delimiting species using DNA and morphological variation and discordant species limits in spiny lizards (*Sceloporus*). *Syst. Biol.* **51**, 69–91 (2002).
- Holdaway, R. N. & Jacomb, C. Rapid extinction of the moas (Aves: Dinornithiformes): model, test, and implications. *Science* **287**, 2250–2254 (2000).
- Polack, J. *New Zealand, Being a Narrative of Travels and Adventures* (Richard Bentley, London, 1838).
- Owen, R. On the bone of an unknown struthious bird from New Zealand, meeting of November 12, 1839. *Proc. Zool. Soc. Lond.* **VII**, 169–171 (1839).
- Worthy, T. H. & Holdaway, R. N. *The Lost World of the Moa: Prehistoric Life of New Zealand* (Canterbury Univ. Press, Christchurch, 2002).
- Clinton, M. & Haines, L. C. An overview of factors influencing sex determination and gonadal development in birds. *Cell. Mol. Life Sci.* **55**, 876–886 (1999).
- Ansari, H., Kakagi, N. & Sasaki, M. Morphological differentiation of sex chromosomes in three species of ratite birds. *Cytogenet. Cell Genet.* **47**, 185–188 (1988).
- Huynen, L., Millar, C. D. & Lambert, D. M. A DNA test to sex ratite birds. *Mol. Ecol.* **11**, 851–856 (2002).
- Griffiths, R., Double, M. C., Orr, K. & Dawson, R. J. A DNA test to sex most birds. *Mol. Ecol.* **7**, 1071–1076 (1998).
- Ellegren, H. First gene on the avian W chromosome (CHD) provides a tag for universal sexing of non-ratite birds. *Proc. R. Soc. Lond. B* **263**, 1635–1641 (1996).
- Taberlet, P. & Luikart, G. Non-invasive genetic sampling and individual identification. *Biol. J. Linn. Soc.* **68**, 41–55 (1999).
- Taberlet, P. *et al.* Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Res.* **24**, 3189–3194 (1996).
- Worthy, T. H. An illustrated key to the main leg bones of moas (Aves: Dinornithiformes). *Nat. Mus. NZ Misc. Ser.* **17**, 1–27 (1988).
- Cooper, A. in *Avian Molecular Evolution and Molecular Systematics* (ed. Mindell, D.) 345–373 (Academic Press, New York, 1997).
- Prange, H. D., Anderson, J. F. & Rahn, H. Scaling of skeletal mass to body mass in birds and mammals. *Am. Nat.* **113**, 103–122 (1979).
- Lambert, D. M. *et al.* Rates of evolution in ancient DNA from Adélie penguins. *Science* **295**, 2270–2273 (2002).
- Sambrook, J., Fritsch, E. F. & Maniatis, T. *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, New York, 1989).
- Griffiths, R. & Tiwari, B. Sex of the last wild Spix's macaw. *Nature* **375**, 454 (1995).
- Swofford, D. L. *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods) Version 4.0b10* (Sinauer Associates, Sunderland, Massachusetts, 2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This research was made possible by grants from the Marsden Fund, Massey University and the Centres of Research Excellence Fund. We are grateful to the following institutions and people: Museum of New Zealand Te Papa Tongarewa, Canterbury Museum, Auckland Institute and Museum, Massey University, Otago Museum, Waitomo Caves Museum, Whanganui Regional Museum, Te Manawa Museum, Institute of Geological and Nuclear Sciences, J. Anderson, J. A. Bartle, A. Tennyson, T. Worthy and V. Ward. We appreciate the support of a number of iwi, especially Ngai Tahu and Ngati Kahungunu.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.M.L. (D.M.Lambert@massey.ac.nz).

True paternal care in a multi-male primate society

Jason C. Buchan¹, Susan C. Alberts^{1,2}, Joan B. Silk³
& Jeanne Altmann^{2,4,5}

¹Department of Biology, Duke University, Box 90338, Durham, North Carolina 27708, USA

²Institute of Primate Research, National Museums of Kenya, Nairobi, Kenya

³Department of Anthropology, UCLA, Los Angeles, California 90095, USA

⁴Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544, USA

⁵Department of Conservation Biology, Brookfield Zoo, Brookfield, Illinois 60513, USA

Although male parental care is rare among mammals¹, adult males of many cercopithecine primate species provide care for infants and juveniles. This care is often in the form of grooming, carrying, support in agonistic interactions, and protection against infanticide^{2,3}. For these behaviours to be interpreted as true parental care, males must selectively direct care towards their own offspring and this care must result in fitness benefits⁴. With the exception of males defending probable offspring from infanticide⁵, male primates living in multi-male, multi-female social groups have not been shown to selectively direct care towards their own offspring^{6,7}. We determined paternity for 75 juveniles in a population of wild savannah baboons (*Papio cynocephalus*) and collected data on interventions in agonistic disputes by adult males on behalf of juveniles as a form of male care. Here we show that adult males differentiate their offspring from unrelated juveniles and selectively support their offspring in agonistic disputes. As support in agonistic disputes is likely to contribute to rank acquisition and protect juveniles from injury and stress^{2,3,5}, this can be considered true parental care.

The subjects of the study were members of five wild savannah baboon groups in Amboseli, Kenya, and adjacent areas at the foot of Mount Kilimanjaro; the study population has been under continuous observation since 1971 (ref. 8). We collected data on interventions in agonistic disputes between July 1999 and July 2002. We also unambiguously identified the fathers for 75 of the 102 juveniles present during this period by analysing six microsatellite loci. The scope and extent of paternal care in baboons could be limited by the fact that males regularly transfer between social groups, so that fathers may leave a group before or shortly after the birth of infants they have sired. However, half of the 75 juveniles in our study still had their fathers present in their social group when the juveniles were 3 years of age, indicating that many males had the opportunity to provide paternal care.

Males helped their own genetic offspring significantly more often than they helped juveniles to whom they were not related (Fig. 1a; Wilcoxon matched-pairs signed-ranks test, $T = 8.5$, $P < 0.01$, $n = 15$ males that had both genetic offspring and unrelated juveniles present). All but 3 of the 15 males who had the opportunity to help genetic offspring and unrelated juveniles provided more care to their own offspring than to unrelated juveniles (Fig. 1a). This pattern of biased care towards offspring will arise if males can distinguish their own offspring from unrelated juveniles. However, it might also arise if males intervene at random in disputes that occur in their proximity, and they happen to be in proximity to their offspring more than expected by chance (if, for instance, they tend to be in proximity to females with whom they have mated in the past). If this 'random intervention' hypothesis for the biasing mechanism is correct, then males would intervene against their offspring as often as they intervene on behalf of their offspring. Our data do not support this 'random intervention' hypothesis. In 73

cases, the intervening male was known to be the father of one participant and unrelated to the other. Males intervened on behalf of their offspring in 69 of these interventions, and against their offspring in only 4 (Wilcoxon matched-pairs signed-ranks test, $T = 3$, $P < 0.008$, $n = 13$ males). The observed pattern of care might also occur if males biased care in favour of juveniles and this bias results in males providing more care to their own offspring by chance. However, if this 'age biased' hypothesis is correct, then males would intervene on behalf of their offspring as often as they intervene on behalf of unrelated juveniles when interactions are

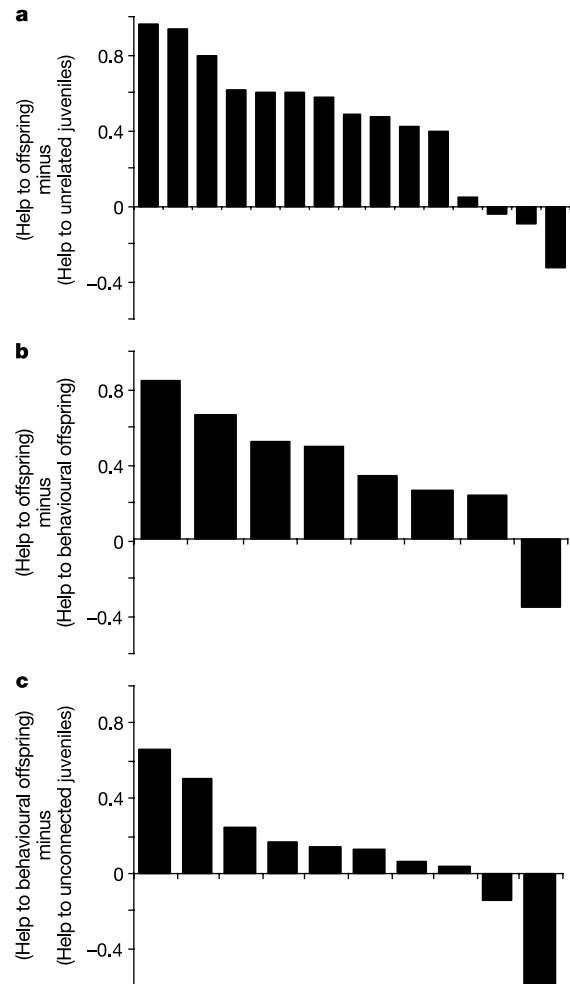


Figure 1 We counted all instances of help that each juvenile received from adult males. We measured the fraction of that help that each male gave to each juvenile, restricting the analysis in each case to the time period in which the male was present in the group. Two-tailed Wilcoxon matched-pairs signed-ranks tests were used to compare the amount of help males provided to juveniles in two different categories. Each bar represents the difference, for one male, between the fraction of help given to juveniles of each type. **a**, Difference, for each male, between the fraction of help he gave to his genetic offspring and to unrelated juveniles (Wilcoxon matched-pairs signed-ranks test, $P < 0.01$). Bars above the expected value of zero represent males that helped genetic offspring more than unrelated juveniles. **b**, Difference, for each male, between the fraction of help he gave to his genetic offspring and to his non-genetic 'behavioural' offspring ($P = 0.05$). Bars above the expected value of zero represent males that helped genetic offspring more than non-genetic behavioural offspring. **c**, Difference, for each male, between the fraction of help he gave to non-genetic behavioural offspring and to 'unconnected' juveniles ($P > 0.1$). Bars above the expected value of zero represent males that helped non-genetic behavioural offspring more than unconnected juveniles.

limited to those between juveniles. We considered the 28 cases in which both participants were juveniles, and the intervening male was known to be the father of one participant and unrelated to the other. Males intervened on behalf of their offspring in 24 of these interventions and against their offspring in only 4 (Wilcoxon matched-pairs signed-ranks test, $T = 4$, $P < 0.009$, $n = 12$ males).

Males were evidently able to distinguish their own offspring from unrelated juveniles. There are two types of mechanism of identification: direct discrimination of kin by using a process such as phenotype matching^{9,10}, or the use of behavioural rules-of-thumb based, for example, on sexual access to fertile females. To address the mechanisms underlying our results, we identified three non-overlapping classes of juveniles for each male. The first class comprised his genetic offspring. The second comprised his non-genetic 'behavioural' offspring: those juveniles for which the male consorted with the mother during the days of most likely conception for that juvenile, but for which he was not the genetic father (see Methods for a description of female sexual cycles and days of likely conception). The final class comprised unrelated juveniles that were neither genetic nor behavioural offspring (apparently 'unconnected' juveniles).

If males use phenotype matching, then we expect males to give more help to genetic than to non-genetic behavioural offspring, and not to differentiate between non-genetic behavioural offspring and unconnected juveniles. In contrast, if males use a simple rule-of-thumb based on whether or not they consorted with the mother when she was likely to conceive, males will not differentiate between genetic and non-genetic behavioural offspring, but will help non-genetic behavioural offspring more often than unconnected juveniles.

Most males supported genetic offspring more than non-genetic behavioural offspring (Wilcoxon matched-pairs signed-ranks test, $T = 4$, $P = 0.05$, $n = 8$ males with both genetic offspring and non-genetic behavioural offspring present; Fig. 1b). In contrast, there was no significant difference in the proportion of care provided to non-genetic behavioural offspring and unconnected juveniles (Wilcoxon matched-pairs signed-ranks test, $T = 13$, $P > 0.10$, $n = 10$ males with both non-genetic behavioural offspring and 'unconnected' juveniles present; Fig. 1c).

Although phenotype matching would explain this result, additional behavioural information, beyond simply whether or not they consorted with the mother during the days of likely conception, might shape males' paternal behaviour. For instance, males might respond to information about what proportion of the mother's 'available consort time' (total observed hours of mate guarding) they monopolized during the days of likely conception. This value was a strong predictor of paternity: the more of the female's consort time that a male monopolized during the days of likely conception, the more likely he was to father a given offspring (logistic regression, χ^2 approximation = 11.52, $P < 0.001$, $n = 63$). Hence this would serve well as a behavioural cue of paternity. Indeed, the probability that a male supported a juvenile at least once during the study was predicted by the proportion of its mother's consort time that he had monopolized during the days of likely conception (logistic regression, χ^2 approximation = 6.57, $P = 0.01$, $n = 43$). However, this relation between mating behaviour and paternal care does not allow us to differentiate between phenotype matching and mating information as the source of kin recognition cues. It only confirms that information about male mating behaviour relative to days of female fertility predicts paternity; males' use of this will be subject to constraints of memory and knowledge about female fertility. On the other hand, if phenotypic markers of relatedness are readily appropriated for kin recognition, then males may use these to differentiate offspring, and the observed relation between mating behaviour and paternal care would still be obtained. Studies in humans suggest a role for olfactory cues in signalling ovulation¹¹, in mate choice^{12–14} and in

providing cues of kin recognition¹⁵. Olfactory cues play well-documented roles in kin recognition in several other mammalian species as well¹⁶. There is also evidence that primates may use visual cues to recognize kin¹⁷. Hence, whether phenotype matching (for example, olfactory or visual cues) or behavioural cues (memory of mating behaviour combined with knowledge about female fertility) or their combined effects represent the more parsimonious explanation for our results is not clear.

A growing body of evidence indicates that primates can and do recognize paternal kin^{5,18–21}. In several cases, evidence indicates that primates rely mainly on behavioural cues, such as age proximity, residence patterns or prior mating behaviour to identify paternal kin. Several of these studies, including two in the Amboseli study population^{19,20}, suggest that phenotype matching may complement behavioural cues in paternal kin recognition^{18–20}. These findings concur with the current study but stand in stark contrast to several earlier studies, which failed to find evidence for paternal kin recognition in the laboratory (reviewed in ref. 22) and which led most researchers to conclude that primates do not recognize paternal kin. We suggest that the discrepancy between results obtained in the field and in the laboratory may result from the fact that animals in the wild rely on multiple cues, whereas laboratory experiments are designed to isolate single cues. In distinguishing offspring, males may use their own mating behaviour in combination with cues of female fertility, phenotype matching and even maternal behaviour towards potential fathers²³. The fact that male baboons in Amboseli selectively support their own offspring indicates that even in species where females mate with multiple males, paternal care can and does evolve. □

Methods

Genetic analysis

Faecal (ref. 24) ($n = 172$ individuals) or blood (ref. 8) ($n = 22$ individuals) samples were collected from 194 baboons, including 79 infants, their mothers and putative fathers. Wherever possible, multiple faecal samples were collected for each individual. Faecal DNA extraction was done by using the QIAamp DNA Stool Mini Kit (QIAGEN GMBH) following the protocol for isolation of DNA from stool for human DNA analysis but with two modifications. We used 500 mg rather than 200 mg of wet faeces, and did not use InhibitEX tablets; instead, 1.2 ml of supernatant from step 3 was divided equally into two 2 ml microcentrifuge tubes containing 25 μ l of Proteinase K as in step 9 of the QIAGEN protocol. Each tube was then carried through as per the protocol with the lysate from both tubes filtered through the same spin column. DNA samples were eluted in 200 μ l of buffer AE. The amount of amplifiable DNA in all samples was quantified by using the 5' nuclease assay as described in ref. 25. Four tetra-nucleotide (D4s243, D14s306, D10s611, D5s1457) and two di-nucleotide human microsatellite loci (D7s503, D13s159) were amplified for each sample. All microsatellites were analysed by using an ABI PRISM3700 DNA analyser and Genescan and Genotyper software. The number of replicates necessary to ensure the detection of allelic dropout was calculated based on the amount of DNA per reaction and the observed rates of allelic dropout^{25,26}.

Maternity for all offspring was known from observations of pregnancies and observations at or shortly after parturition. To ensure the correct identity of samples, mothers and offspring were checked for mendelian mismatches, and all putative fathers were genotyped from two separate faecal samples. If mismatches occurred, DNA was extracted from further faecal samples until the mismatch was resolved. All males that were at least four years of age and present during the conception of an infant were considered putative fathers, although males do not reach full adulthood until approximately 7.5 years (ref. 27).

Initially, duplicate polymerase chain reactions (PCRs) of all samples were done. If both PCRs produced identical heterozygotes, the genotype was considered to be correct. If, after two PCRs, an individual's genotype consisted of both a heterozygote and homozygote with a common allele, further replications were done until both alleles were observed at least twice. Individuals whose genotypes appeared to be homozygous were amplified seven times if samples contained less than 200 μ g μ l⁻¹ and four times for samples with more than 200 μ g μ l⁻¹ of DNA. Paternity was based on exclusion and further supported through the use of the likelihood-based paternity assignment program CERVUS 2.0 (ref. 28).

Behavioural data

Savannah baboons mate in the context of mate-guarding episodes that occur during the follicular phase of the female's cycle, when she has a prominent swelling of the sex skin²⁹. Data on the identity of male and female partners in all mate-guarding episodes were collected as part of regular monitoring of study groups. We identified behavioural offspring as those offspring for which the male was observed to mate with the mother during the most likely days of conception for that juvenile, that is, during the last five days of the follicular phase of the sexual cycle²⁹. The end of the follicular phase is signalled by deflation of the swollen sexual skin; the days of most likely conception are identified

retrospectively based on near-daily records of swelling size and state. Observers recorded all agonistic interactions and interventions on an *ad libitum* basis. In each case of agonism, observers recorded the identity of individuals involved in the aggressive encounter and its outcome³⁰. When third parties intervened in disputes, observers recorded the identity of the individual who intervened (ally), the identity of the individual that received support, the identity of the individual against whom support was directed, and the type of support that was provided. Support took two forms. Allies either directed overt aggression towards one of the participants (designated the opponent), or established close proximity or affiliative physical contact with one of the participants (designated the beneficiary). See ref. 31 for more details of data collection; methods employed in that study were identical to those employed here.

Adult males supported juveniles in 193 disputes. In 93 of these events, both disputants were juveniles (less than 4 years old); in 36, the opponent was an adult female; in 49, a subadult male; and in 15, an adult male (see ref. 27 for definitions of subadult and adult males). The median age of juvenile recipients of adult male help was 2.25 yr; the interquartile range was 1.75 yr to 3.13 yr, and the youngest recipient was 4 months old.

Received 15 April; accepted 30 June 2003; doi:10.1038/nature01866.

- Woodroffe, R. & Vincent, A. Mother's little helpers: patterns of male care in mammals. *Trends Ecol. Evol.* **9**, 294–297 (1994).
- Smuts, B. B. & Gubernick, D. J. In *Father–Child Relations* (ed. Hewlett, B. S.) 1–30 (Aldine De Gruyter, New York, 1992).
- van Schaik, C. P. & Paul, A. Male care in primates: does it ever reflect paternity? *Evol. Anthropol.* **5**, 152–156 (1996).
- Trivers, R. L. In *Sexual Selection and the Descent of Man 1871–1971* (ed. Campbell, B.) (Aldine, Chicago, 1972).
- Borries, C., Launhardt, K., Epplen, C., Epplen, J. T. & Winkler, P. Males as infant protectors in Hanuman langurs (*Presbytis entellus*) living in multimale groups—defence pattern, paternity and sexual behaviour. *Behav. Ecol. Sociobiol.* **46**, 350–356 (1999).
- Paul, A., Kuester, J. & Arnemann, J. The sociobiology of male–infant interactions in Barbary macaques. *Macaca sylvanus. Anim. Behav.* **51**, 155–170 (1996).
- Ménard, N. *et al.* Is male–infant caretaking related to paternity and/or mating activities in wild Barbary macaques (*Macaca sylvanus*)? *C.R. Acad. Sci. III Vie* **324**, 601–610 (2001).
- Altmann, J. *et al.* Behavior predicts genetic structure in a wild primate group. *Proc. Natl Acad. Sci. USA* **93**, 5797–5801 (1996).
- Sherman, P. W., Reeve, H. K. & Pfennig, D. W. In *Behavioural Ecology: An Evolutionary Approach* (eds Krebs, J. R. & Davies, N. B.) 69–96 (Blackwell Scientific, Oxford, 1997).
- Hauber, M. E. & Sherman, P. W. Self-referent phenotype matching: theoretical considerations and empirical evidence. *Trends Neurosci.* **24**, 609–616 (2001).
- Singh, D. & Bronstad, P. Female body odour is a potential cue to ovulation. *Proc. R. Soc. Lond. B* **286**, 797–801 (2001).
- Wedekind, C., Seebeck, T., Bettens, F. & Paepke, A. J. MHC-dependent mate preferences in humans. *Proc. R. Soc. Lond. B* **260**, 245–249 (1995).
- Wedekind, C. & Furi, S. Body odour preferences in men and women: do they aim for specific MHC combinations or simply heterozygosity? *Proc. R. Soc. Lond. B* **264**, 1471–1479 (1996).
- Ober, C. *et al.* HLA and mate choice in humans. *Am. J. Hum. Genet.* **61**, 497–504 (1997).
- Jacob, S., McClintock, M. K., Zelano, B. & Ober, C. Paternally inherited HLA alleles are associated with women's choice of male odours. *Nature Genet.* **30**, 175–179 (2002).
- Halpin, Z. T. In *Kin Recognition* (ed. Hepper, P. G.) 220–258 (Cambridge Univ. Press, Cambridge, 1991).
- Parr, L. A. & de Waal, F. B. M. Visual kin recognition in chimpanzees. *Nature* **399**, 647–648 (1999).
- Widdig, A., Nürnberg, P., Krawczak, M. & Bercovitch, F. B. Paternal relatedness and age proximity regulate social relationships among adult female rhesus macaques. *Proc. Natl Acad. Sci. USA* **98**, 13769–13773 (2001).
- Smith, K., Alberts, S. C. & Altmann, J. Wild female baboons bias their social behaviour towards paternal half-sisters. *Proc. R. Soc. Lond. B* **270**, 503–510 (2003).
- Alberts, S. C. Paternal kin discrimination in wild baboons. *Proc. R. Soc. Lond. B* **266**, 1501–1506 (1999).
- Soltis, J., Thomsen, R., Matsubayashi, K. & Takenaka, O. Infanticide by resident males and female counter-strategies in wild Japanese macaques (*Macaca fuscata*). *Behav. Ecol. Sociobiol.* **48**, 195–202 (2000).
- Silk, J. B. Kin selection in primate groups. *Int. J. Primatol.* **23**, 849–875 (2002).
- Palombit, R. A., Seyfarth, R. M. & Cheney, D. L. The adaptive value of 'friendships' to female baboons: experimental and observational evidence. *Anim. Behav.* **54**, 599–614 (1997).
- Smith, K. L. *et al.* Cross-species amplification, non-invasive genotyping, and non-mendelian inheritance of human STRPs in savannah baboons. *Am. J. Primatol.* **51**, 219–227 (2000).
- Morin, P. A., Chambers, K. E., Boesch, C. & Vigilant, L. Quantitative polymerase chain reaction analysis of DNA from noninvasive samples for accurate microsatellite genotyping of wild chimpanzees (*Pan troglodytes verus*). *Mol. Ecol.* **10**, 1835–1844 (2001).
- Taberlet, P. *et al.* Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Res.* **24**, 3189–3194 (1996).
- Alberts, S. C. & Altmann, J. Preparation and activation: determinants of age at reproductive maturity in male baboons. *Behav. Ecol. Sociobiol.* **36**, 397–406 (1995).
- Marshall, T. C., Slate, J., Kruuk, L. E. B. & Pemberton, J. M. Statistical confidence for likelihood-based paternity inference in natural populations. *Mol. Ecol.* **7**, 639–655 (1998).
- Shaikh, A. A., Celaya, C. L., Gomez, I. & Shaikh, S. A. Temporal relationship of hormonal peaks to ovulation and sex skin deturgescence in the baboon. *Primates* **23**, 444–452 (1982).
- Hausfater, G. *Dominance and Reproduction in Baboons (Papio cynocephalus)* (Basel, Karger, 1975).
- Silk, J. B., Alberts, S. C. & Altmann, J. Patterns of coalition formation by adult female baboons in Amboseli, Kenya. *Anim. Behav.* (in the press).

Acknowledgements We thank the Office of the President of the Republic of Kenya and the Kenya Wildlife Service for permission to work in Amboseli, and the Institute of Primate Research for local sponsorship. We also thank the Wardens and staff of Amboseli National Park, and the pastoralist communities of Amboseli and Longido for continuous cooperation and assistance. We acknowledge financial support from the National Science Foundation (to J.A., J.B.S. and S.C.A.), the Chicago Zoological Society (to J.A.), the L.S.B. Leakey Foundation and the National Geographic Society (J.B.S.). We thank M. Lavine, of the Duke University Statistical Consulting Center, and R. Zimmerman for advice and assistance with statistical tests; R. S. Mututua, S. N. Sayialel and J. K. Warutere for data and sample collection. C. Packer and R. Palombit provided comments on the manuscript.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.C.A. (alberts@duke.edu).

Motion-induced spatial conflict

Derek H. Arnold & Alan Johnston

Department of Psychology and Institute of Cognitive Neuroscience, University College London, Gower Street, London WC1E 6BT, UK

Borders defined by small changes in brightness (luminance contrast) or by differences in colour (chromatic contrast) appear to move more slowly than those defined by strong luminance contrast^{1–4}. As spatial coding is influenced by motion^{5–7}, if placed in close proximity, the different types of moving border might appear to drift apart⁸. Using this configuration, we show here that observers instead report a clear illusory spatial jitter of the low-luminance-contrast boundary. This visible interaction between motion and spatial-position coding occurred at a characteristic rate (~22.3 Hz), although the stimulus motion was continuous and invariant. The jitter rate did not vary with the speed of movement. The jitter was not due to small involuntary movements of the eyes, because it only occurred at a specific point within the stimulus, the low-luminance-contrast boundary. These findings show that the human visual system contains a neural mechanism that periodically resolves the spatial conflict created by adjacent moving borders that have the same physical but different perceptual speeds.

A bright red dot moving against a dark background provides a strong luminance-defined motion signal. A smaller equiluminant green dot superimposed on this target provides a weaker motion signal at the chromatic boundary. To the extent that motion influences spatial position^{5–7}, the green dot might be expected to lag progressively behind. This scheme has recently been suggested as an explanation for the classical 'fluttering hearts' illusion⁸ (Fig. 1).

When we created this configuration (Fig. 2a), it was clear that the two parts of the stimulus did not appear to drift apart. However, a vivid perceptual illusion was immediately apparent. When fixation was maintained on a stationary target, the spatial position of the green dot appeared to jitter while moving. To examine this phenomenon, we constructed a stimulus consisting of four dots (Fig. 2a). A small green dot was superimposed on a larger red dot to form a bull's-eye configuration. Another green dot, of the same size, was shown against a dark background (isolated motion). All these dots rotated about a central static fixation point at a constant retinal velocity of $6.75^\circ \text{ s}^{-1}$. During a run of trials, we systematically manipulated the luminance of the green dots. On each trial, observers were required to indicate whether the foreground green dot or, in different trial runs, the isolated green dot appeared to jitter while moving. Jitter was reported most often when there was little or no luminance contrast between the moving foreground and back-

ground components (Fig. 3). None of our observers ever saw jitter in the isolated-motion condition.

To determine whether the perceived jitter is a general property of equiluminant motion, we presented a rotating green dot against a static red ring (Fig. 2b). By varying the luminance of the green dot, in separate trials, we could test for positional jitter at a range of luminance contrasts spanning the equiluminant point. None of our observers reported jitter for any luminance difference between the green dot and the background (Fig. 3). As jitter does not occur for isolated low-luminance-contrast motion, it is clear that the illusory jitter requires an interaction between signals generated by high and low luminance contrast (or chromatic borders). Because the effect is not generated by chromatic equiluminant motion *per se*, the illusory jitter cannot be caused by a difference in the speed with which the visual system responds to different colours⁹, or by inhibitory interactions between rod- and cone-mediated signals at the colour boundary¹⁰. In any case, although processing delays may introduce a perceptual lag, they cannot provide a sufficient explanation for positional jitter (Fig. 1).

If a difference in perceived speed between low- and high-luminance-contrast boundaries is the critical determining factor, rather than some distinguishing feature of colour processing, then jitter should also be visible for achromatic borders. We constructed a stimulus that contained two green squares presented 2° above and below a central static fixation point (Fig. 2c). The luminance of a central region within the squares was either incremented or decremented from trial to trial to generate a low-luminance-contrast boundary. The squares moved in counter-phase from left to right at a retinal velocity of 4.6° s⁻¹ with a periodicity of 0.5 Hz. Observers reported illusory jitter only when there was a small luminance contrast between the foreground and background regions of the target (Fig. 3).

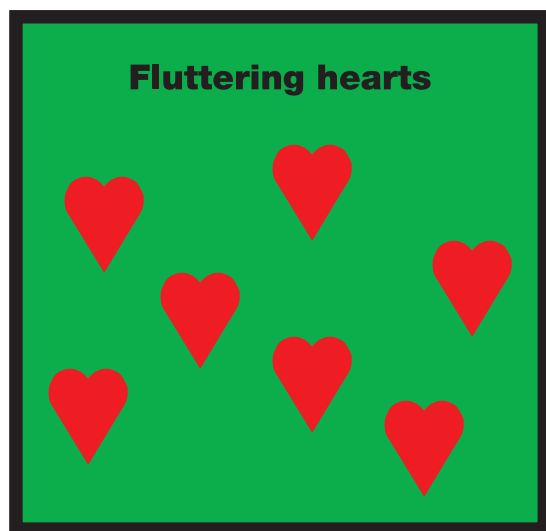


Figure 1 'Fluttering hearts' illusion. If you shake this page from side to side, the red hearts may seem to shift in position relative to the black frame. The figure combines borders of high (dark frame and surround) and low (red hearts and surround) luminance contrast. Originally, the illusion was thought to reflect differences in the speed of visual processing for different colours^{9,10}. More recently, the illusion has been attributed to a difference in the perceived speeds of the two types of moving border⁸. If the stimulus motion is kept constant, in the absence of a compensatory mechanism, the first hypothesis posits that the perceived positions of the hearts should lag that of the frame by a fixed interval. The more recent hypothesis suggests that the borders should drift apart. Here, using a configuration containing adjacent high- and low-luminance-contrast borders, we show that the low-contrast border appears to jitter. This is consistent with a corrective process that periodically snaps the perceptually slower-moving border back into spatial alignment with the faster-moving border (see Supplementary Information for demonstrations).

The illusory jitter could result from small eye movements that are present even during target fixation^{11–13}. To investigate this, we used two further configurations. In one, observers tracked the centre of the bull's-eye as it rotated (Fig. 2d). In this situation, none of the observers ever reported illusory jitter (Fig. 3). In the second, observers tracked a rotating fixation point that had the same speed and eccentricity as the target in the preceding stimulus configurations. A static red–green bull's-eye replaced the fixation point (Fig. 2e). Again, on different trials, we systematically manipulated the luminance of the central green dot while we kept the luminance of the red background dot constant. In this condition, robust illusory jitter of the foreground dot was clearly visible (Fig. 3).

These findings show that the illusory jitter is dependent on retinal and not physical motion of the bull's-eye. When the retinal position of the bull's-eye is kept relatively constant, by tracking (Fig. 2d), there is no illusory jitter (Fig. 3). We also found that there is no illusory jitter of an extrafoveal bull's-eye that moves in phase with a tracked bull's-eye target, demonstrating that the elimination of jitter for tracked targets is not simply a result of foveation.

Adding an eye-velocity vector, extracted from a copy of the motor commands or from proprioceptive extra-retinal signals, to retinal velocities while tracking could stabilize the position of static patterns^{11–13}. However, as this signal would provide a universal image compensation, it should affect all borders equally and could not explain the spatially specific positional jitter of the static bull's-eye (Fig. 2e). During informal observations, we have also found that illusory jitter can be seen when the stimulus is viewed monocularly through a pin-hole, ruling out accommodative micro-fluctuations¹⁴ as a causal mechanism, and when presentation is restricted to just 100 ms, a time span that is too short to allow saccadic movements of the eye.

Illusory jitter in a static pattern has been reported after adaptation to physical jitter in an adjacent part of the visual field^{15–16}. This illusion has been taken as evidence for a mechanism that stabilizes the visual image by calculating a baseline velocity from the region of the retina that has the lowest instantaneous velocity, and subtracting this baseline from all image velocities^{15–16}. The illusory jitter reported here differs from the jitter after-effect because the after-effect results in synchronous jitter for all non-adapted locations, whereas the illusory jitter described here is highly specific. While the centre of the bull's-eye is perceived to jitter, the surround

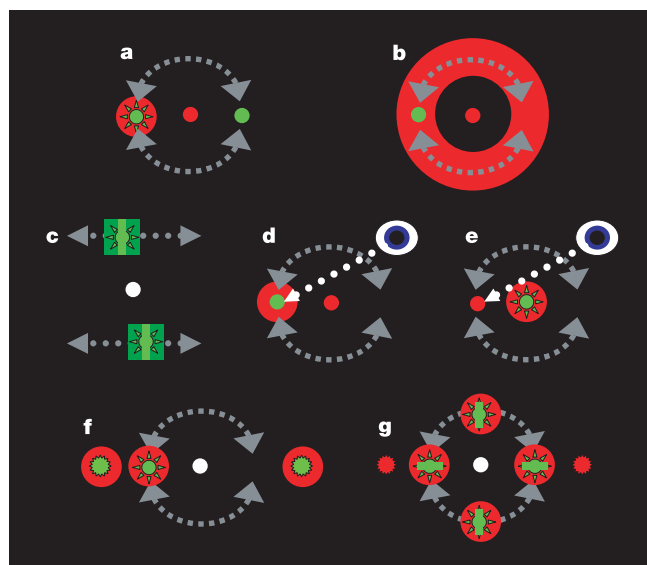


Figure 2 Schematic diagrams of the stimulus configurations. Conditions in which illusory jitter was seen are represented by starburst patterns (a, c, e–g); those in which physical flicker was shown are represented by jagged circles (f, g).

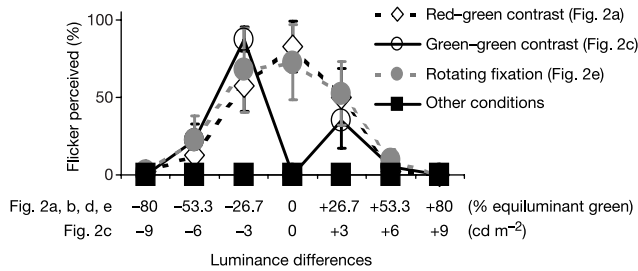


Figure 3 Perceived jitter. Percentage of times that illusory jitter was seen by four observers, naive as to the purpose of the study, as a function of the luminance difference between foreground and background regions of retinally moving stimuli (Fig. 2a, c–e) or between a rotating green dot and a red ring (Fig. 2b). For stimulus configurations shown in Fig. 2a, b, d, e, foreground luminance is expressed as a percentage ($\pm 80\%$) of green luminance perceived to match a red luminance of 18.1 cd m^{-2} , measured separately for each observer by the minimal motion method²⁰. For the configuration shown in Fig. 2c, the luminance of the central green bar was varied within a range $\pm 8.7 \text{ cd m}^{-2}$ from physical equiluminance (18.1 cd m^{-2}). Error bars show $\pm 1 \text{ s.e.m.}$

of the bull's-eye, the isolated motion and the fixation point all appear to be stable (Fig. 2a, e, f).

The illusory jitter appeared to have a characteristic rate. To estimate this rate, we constructed two stimuli in which rates of physical flicker could be matched to the perceived rate of the illusory jitter. The first consisted of three red–green bull's-eyes presented on a cathode-ray tube (CRT) monitor. One rotated around a central fixation point. The other two were static and were presented more eccentrically to the left and right of fixation (Fig. 2f). The luminance of the green centre of the rotating bull's-eye was varied. If illusory

jitter was seen, the observer introduced a sine-wave-profiled luminance flicker to the static green dots and adjusted the rate of this flicker (in steps of 5 Hz by pressing one of two response levers) until it appeared to match the rate of the illusory jitter.

To ensure that our observations were not an artefact resulting from the CRT presentation or from artificial lighting, we printed physically equiluminant red–green stimuli on a black background and rotated this image at a constant speed using an electric motor (Fig. 2g). Robust illusory jitter could be seen when this stimulus was viewed under natural daylight illumination. To estimate the rate of this jitter, we placed two light-emitting diodes (LEDs) eccentrically to the left and right of the stimulus. A pulse generator was used to create a sine-wave-profiled luminance flicker of the LEDs, and observers adjusted this rate until it seemed to match that of the illusory jitter.

Whenever illusory jitter was seen, it was reported to have a constant rate that did not vary as a function of stimulus velocity for any of our observers (Fig. 4a, c, e, g). With CRT presentation, the rate was estimated as being $\sim 25 \text{ Hz}$ (Fig. 4a, c, e). This rate did not appear to vary when the monitor refresh rate was changed (from 100 to 60 Hz). Also, from the results obtained using the CRT presentation, it was evident that the illusory jitter was seen most often when the red–green bull's-eye was at, or close to, subjective equiluminance (Fig. 4b, d, f). However, for faster and slower motions, the illusory jitter was less salient in that it was typically seen less often and over a reduced range of luminance contrasts. With the physical stimulus, the rate of illusory jitter was estimated to be $\sim 22.2 \text{ Hz}$ (Fig. 4g). If the green bars within the physical stimulus were replaced with black, no illusory jitter was ever reported.

The illusory spatial jitter has a characteristic periodicity (~ 22.2 – 25 Hz) that does not seem to vary as a function of the physical speed of the stimulus. This positional jitter cannot readily be attributed to

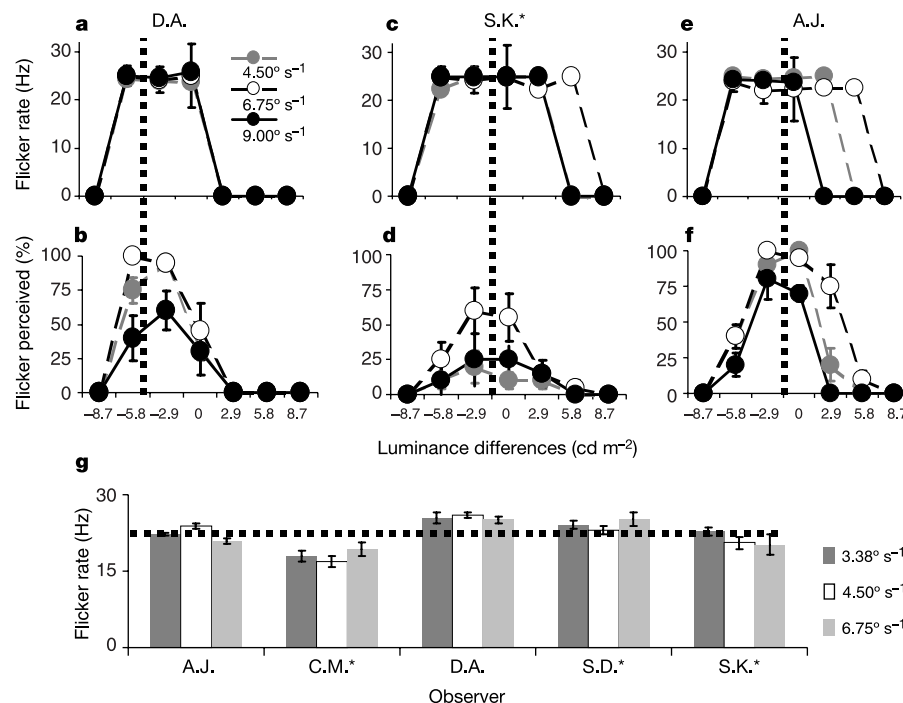


Figure 4 Perceived jitter rate. Rate (a, c, e) and percentage of times (b, d, f) that illusory jitter was seen by three observers as a function of the luminance difference between foreground (9.4 – 26.8 cd m^{-2}) and background (18.1 cd m^{-2}) regions of rotating bull's-eyes. Data for three velocities are shown. Vertical dotted lines show the foreground luminance subjectively equiluminant with the background as determined by a minimum-

motion task²⁰. g, Rate of illusory jitter within a physical sunlit stimulus for five observers as a function of stimulus velocity. The dotted horizontal line depicts the rate of perceived jitter (22.3 Hz) averaged across subjects and stimulus velocities. Error bars show $\pm 1 \text{ s.e.m.}$; asterisks placed next to the observers' initials indicate that they were naive as to the purpose of the study.

the stimulus, which is invariant in its form and motion, or to random eye movements. This leaves the possibility that it reflects a dynamic neural process. When combinations of high- and low-luminance-contrast motion are shown together, as in the examples described here, the motion cues are consistently at variance with the spatial configuration. Unless some form of resolution occurs, the two boundaries might appear to disengage⁸. We propose that the illusory jitter is a visible consequence of this resolution. It has been suggested that reciprocal feedback between, and lateral interactions within, cortical areas can cause synchronous neural spiking with a frequency in the gamma range (20–50 Hz)^{17–19}. The characteristic frequency of the illusory jitter described here might similarly reflect the temporal dynamics of recurrent neural processes that mediate the integration of motion-based spatial predictions and subsequent spatial processing. □

Methods

All stimuli used in the conditions represented in Fig. 2a–f were displayed on a 19-inch Sony Trinitron Multiscan 400PS monitor, with a refresh rate of 100 Hz, driven by a VSG 2/5 visual stimulus generator (Cambridge Research Systems). The standard configuration consisted of a large red dot (Commission Internationale d'Éclairage (CIE) 1931 chromaticity chart: $x = 0.60$, $y = 0.34$) with a diameter subtending 1.5° of visual angle, and a smaller superimposed green dot (CIE 1931: $x = 0.28$, $y = 0.595$) with a diameter subtending 0.5° . In the configurations represented in Fig. 2a, b, d–f, the rotating peripheral bull's-eyes were centred 2.25° of visual angle away from a central fixation point, and in Fig. 2f the additional locations were 3.75° eccentric. In the configuration represented in Fig. 2c, two green squares with a width and height of 1.4° were centred 2° above and below a central static fixation point. The central region had a height of 1.4° and a width of 0.25° . These stimuli were viewed in the dark from a distance of 57 cm with the head placed in a chin rest.

For all configurations, other than those depicted in Fig. 2c, g, the physical direction of motion could be clockwise or anti-clockwise, determined at random from trial to trial. During each trial in conditions Fig. 2a–f, the stimulus remained until the observer reported whether jitter was visible or not by pressing one of two levers. In these conditions, during a run of trials, seven luminance levels of the target stimulus were sampled ten times. Each data point in Fig. 3 and in Fig. 4a–f is the mean of four runs.

In the flicker-matching experiment illustrated in Fig. 2f, observers adjusted the luminance flicker frequency of peripheral dots in 5 Hz steps. Note that this sine-wave luminance function was sampled at the monitor refresh rate, 100 Hz. The physical stimulus depicted in Fig. 2g contained four red dots, with a diameter subtending 2° centred 2.25° of visual angle away from a central fixation point. Equiluminant green bars, with a height of 1.25° and a width of 0.25° , were centred within the red dots. The direction of rotation was clockwise and the orientation of the bars was orthogonal to the direction of rotation. LEDs were placed 3.75° eccentrically to the left and right of the central fixation point. Before the peripheral LEDs were shown, it was confirmed that each observer could see illusory jitter of the green bars at each of three physical speeds of rotation. Observers adjusted the rate of sine-wave flicker of the LEDs by adjusting an analogue control on a pulse generator until the rate of flicker seemed to match the rate of the illusory spatial jitter. This was done ten times for each of three stimulus velocities by each observer.

Received 3 July; accepted 24 July 2003; doi:10.1038/nature01955.

1. Anstis, S. Footsteps and inchworms: Illusions show that contrast affects apparent speed. *Perception* **30**, 785–794 (2001).
2. Blakemore, M. R. & Snowden, R. J. The effect of contrast upon perceived speed: a general phenomenon? *Perception* **28**, 33–48 (1999).
3. Cavanagh, P., Tyler, C. W. & Favreau, O. E. Perceived velocity of moving chromatic gratings. *J. Opt. Soc. Am. A* **1**, 893–899 (1984).
4. Thompson, P. Perceived rate of movement depends on contrast. *Vision Res.* **22**, 377–380 (1982).
5. De Valois, R. L. & De Valois, K. K. Vernier acuity with stationary moving Gabors. *Vision Res.* **31**, 1619–1626 (1991).
6. Nishida, S. & Johnston, A. Influence of motion signals on the perceived position of spatial pattern. *Nature* **397**, 610–612 (1999).
7. Whitney, D. & Cavanagh, P. Motion distorts visual space: shifting the perceived position of remote stationary objects. *Nature Neurosci.* **3**, 954–959 (2000).
8. Nguyen-Tri, D. & Faubert, J. The fluttering-heart illusion: a new hypothesis. *Perception* **32**, 627–634 (2003).
9. Helmholtz, H. *Treatise on Physiological Optics* (Dover, New York, 1962).
10. von Grunau, M. W. The "fluttering heart" and spatio-temporal characteristics of color processing III. Interactions between the systems of the rods and the long-wavelength cones. *Vision Res.* **16**, 397–401 (1976).
11. Bridgeman, B. & Stark, L. Ocular proprioception and efference copy in registering visual direction. *Vision Res.* **31**, 1903–1913 (1991).
12. Sherrington, C. S. Observations on the sensual role of the proprioceptive nerve supply of the extrinsic ocular muscles. *Brain* **41**, 332–343 (1918).
13. Wiesel, T. N. & Hubel, D. H. Spatial and chromatic interactions in the lateral geniculate body of the rhesus monkey. *J. Neurophysiol.* **29**, 1115–1116 (1966).
14. Campbell, F. W., Robson, J. G. & Westheimer, G. Fluctuations of accommodation under steady viewing. *J. Physiol. (Lond.)* **145**, 579–594 (1959).

15. Murakami, I. & Cavanagh, P. A jitter after-effect reveals motion-based stabilization of vision. *Nature* **395**, 798–801 (1998).
16. Murakami, I. Illusory jitter in a static stimulus surrounded by a synchronously flickering pattern. *Vision Res.* **43**, 957–969 (2003).
17. Von der Marlsburg, C. & Schneider, W. A neural cocktail-party processor. *Biol. Cybern.* **54**, 29–40 (1986).
18. Engel, A. K., König, P., Kreiter, A. K. & Singer, W. Interhemispheric synchronization of oscillatory neuronal responses in cat visual cortex. *Science* **252**, 1177–1179 (1991).
19. Engel, A. K. & Singer, W. Temporal binding and the neural correlates of sensory awareness. *Trends Cogn. Sci.* **5**, 16–25 (2001).
20. Anstis, S. M. & Cavanagh, P. in *Color Vision: Physiology and Psychophysics* (eds Mollon, J. D. & Sharpe, L. T.) 155–166 (Academic, London, 1983).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We are grateful to C. Clifford, J. Dale, F. Kandil, S. Nishida and Q. Zaidi for their suggestions and comments.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.H.A. (derek.arnold@ucl.ac.uk).

Cellular networks underlying human spatial navigation

Arne D. Ekstrom¹, Michael J. Kahana¹, Jeremy B. Caplan¹, Tony A. Fields², Eve A. Isham², Ehren L. Newman¹ & Itzhak Fried^{2,3}

¹Volen Center for Complex Systems, Brandeis University, Waltham, Massachusetts 02454, USA

²Division of Neurosurgery and Department of Psychiatry and Biobehavioral Science, University of California, Los Angeles (UCLA), California 90095, USA

³Functional Neurosurgery Unit, Tel-Aviv Medical Center and Sackler School of Medicine, Tel-Aviv University, Tel-Aviv 69978, Israel

Place cells of the rodent hippocampus constitute one of the most striking examples of a correlation between neuronal activity and complex behaviour in mammals^{1,2}. These cells increase their firing rates when the animal traverses specific regions of its surroundings, providing a context-dependent map of the environment^{3–5}. Neuroimaging studies implicate the hippocampus and the parahippocampal region in human navigation^{6–8}. However, these regions also respond selectively to visual stimuli^{9–13}. It thus remains unclear whether rodent place coding has a homologue in humans or whether human navigation is driven by a different, visually based neural mechanism. We directly recorded from 317 neurons in the human medial temporal and frontal lobes while subjects explored and navigated a virtual town. Here we present evidence for a neural code of human spatial navigation based on cells that respond at specific spatial locations and cells that respond to views of landmarks. The former are present primarily in the hippocampus, and the latter in the parahippocampal region. Cells throughout the frontal and temporal lobes responded to the subjects' navigational goals and to conjunctions of place, goal and view.

Responses of single neurons were recorded in seven subjects who were patients with pharmacologically intractable epilepsy undergoing invasive monitoring with intracranial electrodes to identify the seizure focus for potential surgical treatment (see Methods). Subjects played a taxi driver computer game in which they explored a virtual town, searching for passengers who appeared in random spatial locations and delivering them to fixed target locations (shops, Fig. 1a, b). Before exploring the town, recordings were

made while subjects viewed shop fronts they would later see during the game (Fig. 1c–e). This provided a control for any cellular responses that might be observed based solely on object perception (see Methods).

We recorded from 317 neurons: 67 cells in the hippocampus, 54 in the parahippocampal region, 111 in the amygdala, and 85 in the frontal lobes (see Methods). To determine the nature of cellular responses during spatial navigation, we compared spike rates as a function of the subject's location in the virtual town (place), the object they viewed (view), and their shop or passenger goal. An analysis of variance for each cell across these three factors revealed that 42% of cells responded significantly ($P < 0.05$) to some aspect of the spatial environment, as revealed by a main effect of one or more of the three factors: 26% responded to place, 12% responded to view, and 21% responded to goal. Sixteen per cent of cells showed interaction effects only. To ensure that view responses did not simply reflect perception of objects outside their spatial context, we compared the neural responses to shop fronts viewed prior to navigation. Only 2% of cells (less than the Type I error rate) responded preferentially to specific shop images, suggesting that these responses could not account for the effect of view on firing rate during spatial navigation.

The observation that cells can respond to both place and view raises the question of whether place-responsive cells are in fact coding for place itself, or whether these cells are responding to a subject's view of a given region in our virtual town. The existence of bona fide place cells would require, at a minimum, that these cells do not also respond to view or to conjunctions of place and view. We therefore asked whether the number of cells responding to place but not view were present above the Type I error rate and in what regions these responses were clustered.

Among all cells recorded, 11% fulfilled the criteria for bona fide place selectivity (31 out of 279). Figure 2a illustrates this place selectivity for a cell in the right hippocampus. Place-responsive cells were significantly more prevalent in the hippocampus (24% of cells in the hippocampus were bona fide place responsive cells) than in the frontal lobes, the parahippocampal region and the amygdala ($\chi^2(3) = 11.3$, $P < 0.01$; Fig. 2b). The locations of place fields in place-responsive cells were determined using a spike-shuffling

method to locate regions of high firing rate that exceeded background. Place-responsive cells had a mean of 1.7 non-contiguous place fields, and place fields showed a mean increase in firing rate of 74% compared with the rate outside of the field. As can be seen in representative examples (Fig. 2a and Supplementary Fig. 3), place fields usually occurred in regions that were frequently traversed and showed robust increases in firing rate compared with background.

To determine whether the place responses of our cells were direction-dependent, we compared the normalized firing rate in place fields that were traversed in one direction with the opposite direction across all 33 hippocampal place fields (traversals were selected based on the highest numbers of crossings). The mean of the distribution of firing-rate differences did not differ from zero ($t(32) = 0.32$, $P = 0.70$) and the distribution (Fig. 2c) did not deviate from normality ($\chi^2(9) = 0.88$, $P = 0.99$), suggesting that there was no directional tendency across the population of hippocampal neurons (if the place responses we recorded were unidirectional, the distribution of differences in firing rates would have been different from zero). We further analysed place-responsive neurons to determine whether they were modulated by the subject's

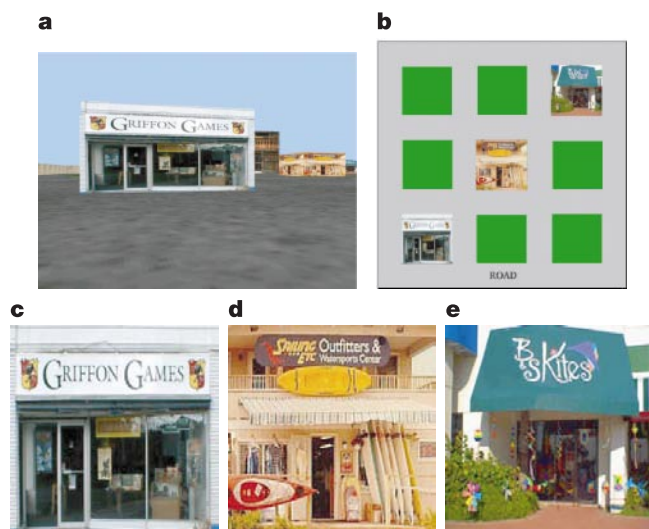


Figure 1 Taxi driver game. **a**, An example of a view seen as a subject navigated through a randomly generated town. Each town contained three labelled, target shops chosen randomly from a pool of 20 possibilities, and 6 unlabelled, non-target buildings chosen from a pool of 48 possibilities. **b**, An example of one particular spatial layout is shown with the corresponding shops (**c–e**) searched for during navigation.

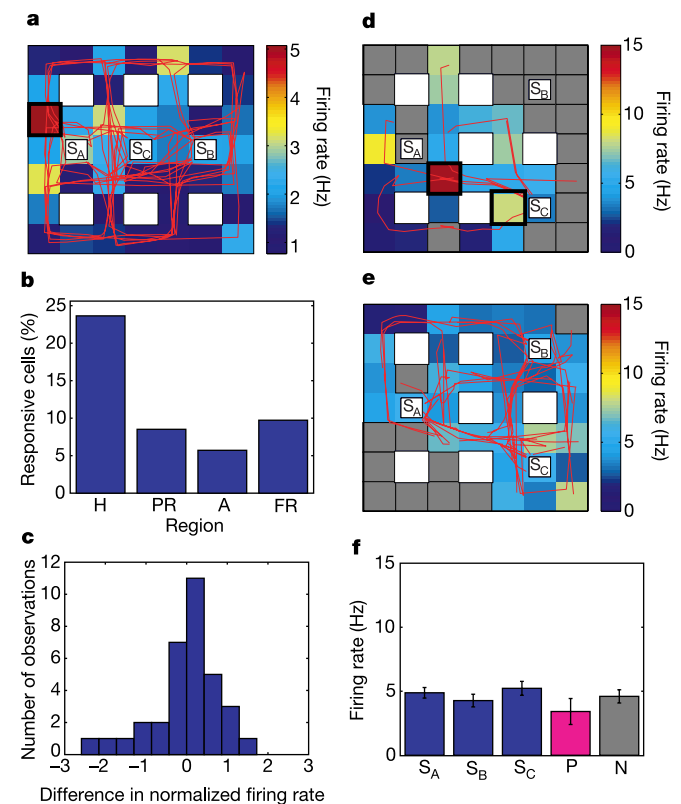


Figure 2 Place-responsive cells. **a**, Firing-rate map of a right hippocampal cell showing significant place selectivity. Letters (S_A, S_B, S_C) indicate shop locations, white boxes indicate non-target buildings, grey boxes indicate unoccupied areas, red lines indicate the subject's trajectory, and black squares indicate regions of significantly high firing rate (all examples, $P < 0.01$; see Methods). **b**, Place-responsive cells were clustered in the hippocampus (H) compared with amygdala (A), parahippocampal region (PR) and frontal lobes (FR). **c**, Regions of high firing included high numbers of traversals in different directions. The distribution of firing-rate differences across these traversals was centred on zero and normal. **d**, Firing-rate map of a right hippocampal cell showing significant place selectivity when searching for shop S_C, but no such specificity when searching for other goals (**e**, areas with < 2 traversals were excluded). **f**, This cell similarly showed no effect of viewing specific targets (P indicates viewing passengers; N indicates a control background view).

goal. Twenty-six per cent of place-responsive cells had place $\times$ goal interaction effects (8 of 31 cells) and fired in different spatial locations depending on the subject's goal; Fig. 2d, e illustrates the response of a goal-modulated place cell recorded from the right hippocampus. When shop S_C was the goal (Fig. 2d), the cell showed clear place-selective responses compared with when shop S_C was not the goal (Fig. 2e). Whereas the cell was strongly modulated by place and goal, it was not modulated by view (Fig. 2f).

Eighty-eight per cent of view cells (29 out of 33) responded preferentially to a single object during navigation (such as a specific shop or passenger, Fig. 3a). Twenty-four view cells were responsive to a specific shop, and among these cells, 14 were location-independent (that is, they showed no place $\times$ view interaction effect and they exhibited a high firing rate in many of the locations where the shop was viewed, Fig. 3b, c). Location-independent view-responsive cells were significantly clustered in the parahippocampal region (Fig. 3f, $\chi^2(3) = 11.3$, $P < 0.01$), where they comprised 7 out of 10 view-responsive cells. Fifteen view cells across anatomical regions (only three of which were in the parahippocampal region) also exhibited place $\times$ view interaction effects. These location-dependent view-responsive cells increased their firing rate when specific shops were viewed from certain spatial locations (Fig. 3d, e).

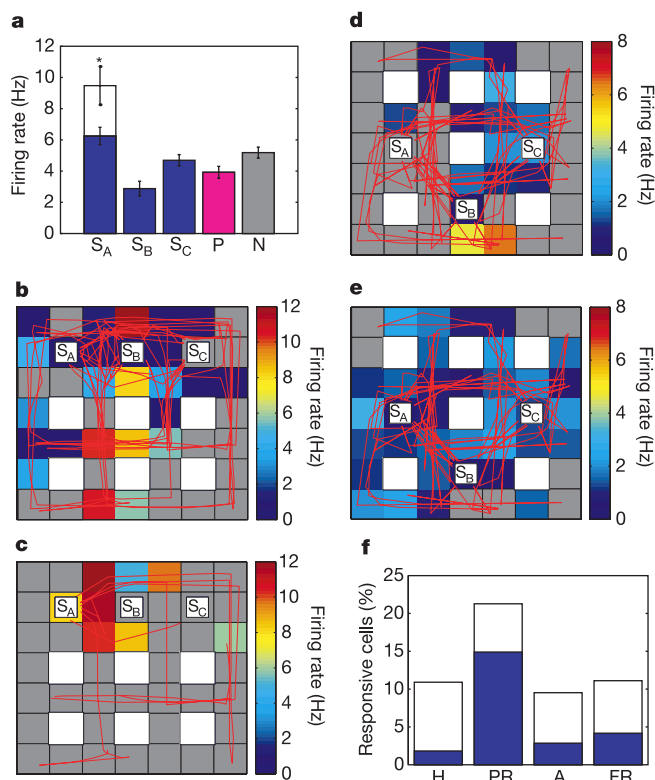


Figure 3 View-responsive cells. **a**, Mean firing rate for a right parahippocampal cell that responded to viewing S_A (as compared with other shops, passengers and control views). The firing rate to viewing S_A (but not other targets) increased significantly when S_A was the goal (white bar). **b**, Firing-rate map shows that this cell responded to viewing S_A from disparate regions; grey regions indicate that S_A was not viewed. **c**, When searching for S_A , the firing rate was consistently high whenever it was viewed. **d**, Firing-rate map of a view-responsive cell in the left amygdala. This cell's activity was modulated by the subject's position; it fired most strongly when S_C was viewed from the town corner nearest to S_B , but not from other spatial positions, and **e**, not when other objects were viewed. **f**, Per cent of location-independent view cells across brain regions. Blue bars, responses to shops; total bar height, responses to all goals (shops and passengers).

Location-dependent view-responsive cells were not clustered by anatomical region.

Twenty-one per cent of cells (59 out of 279) responded to the subjects' goal (that is, one of the target shops (Fig. 4a) or passengers (Fig. 4b)). Although we recorded a smaller percentage of goal-responsive cells in the amygdala than in other regions, this effect was not statistically significant ($\chi^2(3) = 6.7$, $P = 0.1$). Goal cells with no main effects of place fired robustly regardless of spatial position (Supplementary Fig. 2a, b). Fifteen per cent of goal-responsive cells also showed view $\times$ goal interaction effects. These cells increased their firing rate during viewing depending on whether or not the shop was a goal (Fig. 4d and Supplementary Fig. 2d); the majority of these view-dependent goal cells (77%) responded to shops and not to passengers.

The anatomical distribution of place- and view-responsive cells reveal a dissociation between the hippocampus and the parahippocampal region, with the hippocampus specialized for place and the parahippocampal region specialized for view ($\chi^2(1) = 10.5$, $P < 0.005$). This finding, together with functional magnetic resonance imaging (fMRI) studies showing that viewing spatial layouts preferentially activates the parahippocampal region⁹, suggests that the hippocampus and parahippocampal region perform complementary functions during navigation. Although an extensive literature for the rat supports the role of the hippocampus in spatial coding, as do studies in humans⁸, single-unit recordings in primates suggest that the hippocampus responds to spatial views during navigation¹⁴ while the parahippocampal region responds to head direction¹⁵. Because of our experimental design, we are unable to adequately address bearing responses (see Methods), although we note that hippocampal responses to spatial locations have also been observed in primates during virtual and real spatial translocations¹⁶.

The presence of place-goal conjunctive cells in the hippocampus may indicate its role in associating goal-related contextual inputs with place, as has been noted in rats during spatial "remapping"^{4,5,17}.

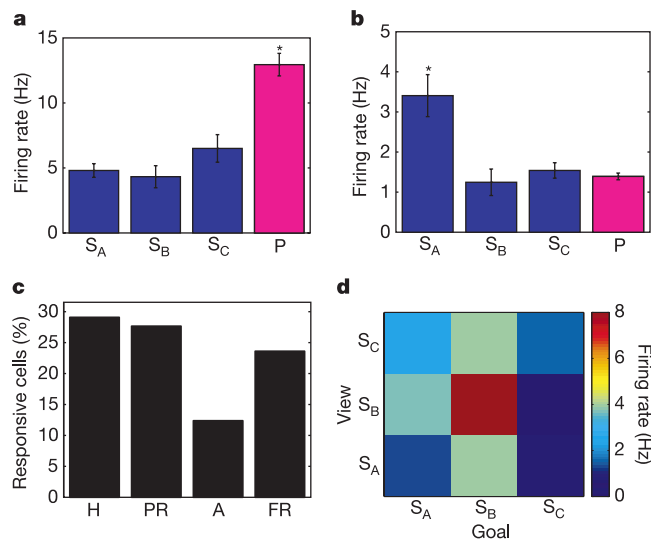


Figure 4 Goal-responsive cells. **a**, Mean firing rate for a right hippocampal cell that responded when seeking passengers (P) and **b**, for a different right hippocampal cell that responded when seeking S_A . **c**, Goal-responsive cells were not significantly clustered by anatomical region. Some goal-responsive cells modulated their firing rate based on what was being viewed, such as this cell in the right amygdala (**d**), which responded preferentially when the goal (shop S_B) was in view. This panel shows firing rates for all combinations of shop being viewed and shop being sought; view $\times$ goal conjunctive cells were not clustered by anatomical region.

Location-dependent and goal-dependent view responses, in contrast, may support navigational strategies that require view-dependent, egocentric representations of space. Goal-dependent view responses may provide information on the progress and success in locating a goal, whereas location-dependent view responses could be useful in planning trajectories to visible goals. We observed some location-dependent view-responsive cells in the parahippocampal region (30% of view cells), although a greater number were location-independent (70% of view cells).

It is intriguing to consider the possibility that projections from the hippocampus to the parahippocampal region may have a role in producing view-independent representations of landmarks in spatial scenes^{9,10}. Our dissociation of parahippocampal and hippocampal function, together with the data discussed above, provide cellular evidence for an emerging model¹⁸ of the physiological basis of human spatial navigation. In this model, the parahippocampal region extracts allocentric spatial information primarily from salient visual landmarks to form a coarse representation of space^{19,20}. The hippocampus combines visual and spatial features, possibly via inputs from the parahippocampal region, with context to compute the flexible map-like representations of space underlying navigation^{21,22}. □

Methods

Behavioral methods

Subjects navigated using the arrow keys on a computer keyboard; when moving, velocity was constant. Virtual towns consisted of six unlabelled, non-target buildings and three labelled, target shops (Fig. 1b). During a single session, subjects made seven deliveries of passengers to each target shop in a random order. Passengers were 'picked up' by driving near them; text then appeared instructing subjects as to which shop the passenger should be delivered. A small box of text in the corner of the screen reminded the subject of their goal. Each delivery began from the random position where the passenger was picked up. Upon delivery of the passenger to a fixed-location shop (accomplished by driving into it), subjects were told whether they had found the correct shop (subjects also received 'virtual' payment for delivering passengers). A text instructed subjects to find another passenger, and subjects explored the city until they located another passenger, at which point the cycle began again. Shops and passengers looked the same from all viewing angles; shops were identified by highly visible names (see ref. 23 for further details concerning the taxi driver game).

Patient data and electrophysiology

Six (of seven) patients were right-handed, two were female; one patient had right temporal-lobe epilepsy, one had left frontal-lobe epilepsy; all others had left temporal-lobe epilepsy (Supplementary Table 1). Each patient had between 6 and 14 depth electrodes implanted bilaterally from a lateral orthogonal approach (surgeries were performed by I.F.). Each of these clinical electrodes terminated with a set of nine 40- μ m platinum-iridium microwires. Signals from these microwires were recorded at 28 kHz and bandpass-filtered between 0.6 and 6 kHz using a 64-channel acquisition system (Neuralynx). Responses of individual cells were isolated based upon the distribution of inter-spike intervals and parameters of the spike waveforms (Supplementary Fig. 1, MClust, developed by A. D. Redish and K. Harris, <http://www.cbc.umn.edu/~redish/mclust>). MRI scans following placement of electrodes, or post-placement computed tomography scans coregistered to preoperative MRI scans, were used to verify the anatomical location of the electrodes (Supplementary Fig. 1a; see also refs 12, 24, 25). All patients provided informed consent. All studies conformed with the guidelines of the Medical Institutional Review Board at UCLA.

The 85 cells in the frontal lobes consisted of cells in anterior cingulate, orbital frontal cortex and in supplementary motor cortex. We use the term 'parahippocampal region', as defined by Witter²⁶, to include pre- and para-subiculum, entorhinal and perirhinal cortices and parahippocampal cortex. Cells with firing rates above 15 Hz were considered to be interneurons and were excluded (6); cells with less than 0.1 Hz firing rate were similarly excluded (32); this left a total of 279 cells for analysis.

Data analysis

Spike counts during different epochs of the taxi driver game were compared using a place (49) $\times$ view (5) $\times$ goal (4) analysis of variance. The place factor could take on one of 49 values representing a 7 $\times$ 7 grid overlaid on each virtual town. The view factor coded for times when subjects viewed shops S_A , S_B or S_C , passenger (P), or background (N). The goal factor coded for times when subjects searched for S_A , S_B , S_C , P (see Supplementary Table 2 for analysis of variance results).

Periods when the subject remained stationary in the game for >500 ms were excluded. Spike-by-position plots were determined by dividing the number of spikes that occurred in a spatial region by the total time spent in that region. Significant 'place fields' were identified by shuffling the spike train randomly and locating firing rates that exceeded 95% of all shuffled spike train firing rates for that region²⁶; all cells identified as place-responsive in our analysis also showed one or more place fields using the spike-shuffling method.

Areas occupied for less than 5 s were not considered, and nor were areas with less than two passes. The view analysis was performed by calculating what the subject was viewing (S_A , S_B , S_C , P, N) every 30 ms. We included only viewing epochs when more than 70% of an object was visible for at least 500 ms; no other objects could be simultaneously visible. The spike train was then restricted to these times to calculate the firing rate while viewing objects during navigation.

To ensure that cellular responses during navigation were not the result of seizure activity, the responses and firing rates of neurons were compared after excluding all cells from areas of seizure focus: this did not affect cellular responses to place, view and goal ($\chi^2(1) = 2.5$, $P > 0.1$), nor firing rate ($\chi^2(1) = 0.1$, $P > 0.1$).

Received 24 March; accepted 17 July 2003; doi:10.1038/nature01964.

- O'Keefe, J. & Dostrovsky, J. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Res.* **34**, 171–175 (1971).
- Muller, R. U., Kubie, J. L. & Ranck, J. B. Spatial firing patterns of hippocampal complex-spike cells in a fixed environment. *J. Neurosci.* **7**, 1935–1950 (1987).
- O'Keefe, J. & Conway, D. H. Hippocampal place units in the freely moving rat: why they fire where they fire. *Exp. Brain Res.* **31**, 573–590 (1978).
- Muller, R. U. & Kubie, J. L. The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells. *J. Neurosci.* **7**, 1951–1968 (1987).
- Wilson, M. A. & McNaughton, B. L. Dynamics of the hippocampal ensemble code for space. *Science* **261**, 1055–1058 (1993).
- Aguirre, G. K., Detre, J. A., Alsop, D. C. & D'Esposito, M. The parahippocampus subserves topographical learning in man. *Cereb. Cortex* **6**, 823–829 (1998).
- Aguirre, G. K., Zarahn, E. & D'Esposito, M. An area within human ventral cortex sensitive to "building" stimuli: evidence and implications. *Neuron* **21**, 373–383 (1996).
- Maguire, E. et al. Knowing where and getting there: a human navigation network. *Science* **280**, 921–924 (1998).
- Epstein, R. & Kanwisher, N. A cortical representation of the local visual environment. *Nature* **392**, 598–601 (1998).
- Epstein, R., Graham, K. & Downing, P. E. Viewpoint specific scene representations in human parahippocampal cortex. *Neuron* **37**, 865–876 (2003).
- Kreiman, G., Koch, C. & Fried, I. Category-specific visual responses of single neurons in the human medial temporal lobe. *Nature Neurosci.* **3**, 946–953 (2000).
- Cameron, K. A., Yashar, S., Wilson, C. L. & Fried, I. Human hippocampal neurons predict how well word pairs will be remembered. *Neuron* **30**, 289–298 (2001).
- Ojemann, G., Schoenfeld-McNeill, J. & Corina, D. Anatomic subdivisions in human temporal cortical neuronal activity related to recent verbal memory. *Nature Neurosci.* **5**, 64–71 (2002).
- Georges-Francois, P., Rolls, E. T. & Robertson, R. G. Spatial view cells in the primate hippocampus: allocentric view not head direction or eye position or place. *Cereb. Cortex* **9**, 197–212 (1999).
- Robertson, R. G., Rolls, E. T. & Georges-Francois, P. Head-direction cells in the primate presubiculum. *Hippocampus* **9**, 206–219 (1999).
- Matsumura, N., Nishijo, H., Tamura, R., Eifuku, S. & Ono, T. Spatial- and task-dependent neuronal responses during real and virtual translocation in the monkey hippocampal formation. *J. Neurosci.* **19**, 2381–2393 (1999).
- Markus, E. J. et al. Interactions between location and task affect the spatial and directional firing of hippocampal neurons. *J. Neurosci.* **15**, 7079–7094 (1995).
- Burgess, N., Maguire, E. & O'Keefe, J. The human hippocampus and spatial and episodic memory. *Neuron* **35**, 625–641 (2002).
- Quirk, G. J., Muller, R. U., Kubie, J. L. & Ranck, J. B. The positional firing properties of medial entorhinal neurons: description and comparison with hippocampal place cells. *J. Neurosci.* **12**, 1945–1963 (1992).
- Frank, L. M., Brown, E. N. & Wilson, M. A. Trajectory encoding in the hippocampus and entorhinal cortex. *Neuron* **27**, 169–178 (2000).
- Tolman, E. C. Cognitive maps in rats and men. *Psych. Rev.* **55**, 189–208 (1948).
- O'Keefe, J. & Nadel, L. *The Hippocampus as a Cognitive Map* (Oxford Univ. Press, Oxford, 1978).
- Caplan, J. B. et al. Human theta oscillations related to sensorimotor integration and spatial learning. *J. Neurosci.* **23**, 4726–4736 (2003).
- Fried, I. et al. Cerebral microdialysis combined with single-neuron and electroencephalographic recording in neurosurgical patients. Technical note. *J. Neurosurg.* **91**, 697–705 (1999).
- Fried, I., MacDonald, K. & Wilson, C. L. Single neuron activity in human hippocampus and amygdala during recognition of faces and objects. *Neuron* **18**, 753–765 (1997).
- Witter, M. in *The Parahippocampal Region: Organization and Role in Cognitive Functions* (eds Witter, M. & Wouterlood, F.) 3–19 (Oxford Univ. Press, Oxford, 2002).
- Skaggs, W. E., McNaughton, B. L., Gothard, K. M. & Markus, E. J. in *Advances in Neural Information Processing Systems* (eds Hanson, S. J., Cowan, J. D. & Giles, C. L.) 1030–1037 (Morgan-Kaufman, San Mateo, California, 1993).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We acknowledge support from NIMH grants to Brandeis University, a NINDS grant to UCLA, and a grant from the Sloan Foundation. We also thank P. Steinmetz, C. Wilson and E. Behnke for technical assistance, and I. Wainwright for editorial assistance.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.J.K. (kahana@brandeis.edu) or I.F. (ifried@mednet.ucla.edu).

Control of spontaneous and damage-induced mutagenesis by SUMO and ubiquitin conjugation

Philipp Stelter & Helle D. Ulrich

Max Planck Institute for Terrestrial Microbiology, Karl-von-Frisch-Strasse, 35043 Marburg, Germany

Protein modification by ubiquitin is emerging as a signal for various biological processes in eukaryotes, including regulated proteolysis, but also for non-degradative functions such as protein localization, DNA repair and regulation of chromatin structure^{1–4}. A small ubiquitin-related modifier (SUMO) uses a similar conjugation system that sometimes counteracts the effects of ubiquitination⁵. Ubiquitin and SUMO compete for modification of proliferating cell nuclear antigen (PCNA), an essential processivity factor for DNA replication and repair⁶. Whereas multi-ubiquitination is mediated by components of the RAD6 pathway and promotes error-free repair, SUMO modification is associated with replication^{6–9}. Here we show that RAD6-mediated mono-ubiquitination of PCNA activates translesion DNA synthesis by the damage-tolerant polymerases η and ζ in yeast. Moreover, polymerase ζ is differentially affected by mono-ubiquitin and SUMO modification of PCNA. Whereas ubiquitination is required for damage-induced mutagenesis, both SUMO and mono-ubiquitin contribute to spontaneous mutagenesis in the absence of DNA damage. Our findings assign a function to SUMO during S phase and demonstrate how ubiquitin and SUMO, by regulating the accuracy of replication and repair, contribute to overall genomic stability.

RAD6-dependent DNA damage tolerance in *Saccharomyces cerevisiae* encompasses an error-free damage avoidance mechanism believed to involve replication fork regression associated with a template switch to the newly synthesized sister chromatid, as well as two pathways of translesion synthesis by polymerase η (Pol η , encoded by RAD30, the yeast homologue of the human xeroderma pigmentosum variant gene XPV) and polymerase ζ (Pol ζ , encoded by REV3 and REV7)^{4,10}. All three branches require the action of the ubiquitin-conjugating enzyme Rad6p in complex with the DNA-binding RING finger protein Rad18p^{11–13}. In the context of error-free bypass the Rad6p–Rad18p complex cooperates with a second ubiquitin-conjugating enzyme, the Ubc13p–Mms2p dimer, and another RING finger protein, Rad5p, in the conjugation of unusual, lysine 63 (K63)-linked multi-ubiquitin chains to K164 of PCNA (POL30)^{6–9}. Consequently, mutation of this residue causes a sensitivity towards DNA-damaging agents that is genetically linked to the error-free branch of the RAD6 pathway⁶. In the absence of Rad5p, Ubc13p or Mms2p PCNA is mono-ubiquitinated by Rad6p and Rad18p⁶, but the question of whether this modification is merely an intermediate on the way to multi-ubiquitin chains or represents a physiologically relevant state of PCNA has not been addressed. SUMO modification of PCNA by Ubc9p and the SUMO-specific ligase Siz1p¹⁴ occurs primarily at K164 as well, but also at K127 (ref. 6). As SUMO modification is most prominent during S phase, it has been suggested to confer a replicative function to PCNA⁶. An inhibitory function of SUMO on repair was inferred from the notion that *pol30(K127/164R)*, a mutant lacking both lysine residues subject to SUMO modification, is more tolerant to DNA damage than *pol30(K164R)*, which is devoid of ubiquitination but can still be modified by SUMO on K127 (ref. 6). Both the absence of modifications in mutants of the conjugation systems as well as the identities of the acceptor lysines on PCNA were verified as previously described⁶ (see Supplementary Fig. 1).

In order to develop a tool for the analysis of ubiquitination of wild-type PCNA in the absence of interfering SUMO modification we examined the effect of a *siz1* deletion on the PCNA lysine mutants. In the *siz1* background both the single- and the double-lysine mutant exhibited virtually identical sensitivities towards ultraviolet irradiation or chemically induced DNA damage, similar to that of *pol30(K127/164R)* in *SIZ1* (Fig. 1a, f), indicating that abolishment of SUMO conjugation has the same effect as the K127R mutation. Thus, the enhanced sensitivity of *pol30(K164R)* compared with *pol30(K127/164R)* in *SIZ1* cells can indeed be attributed to SUMO modification of PCNA at K127. Considering this inhibitory effect of SUMO, we predicted that deletion of *SIZ1*, which by itself has very little influence on repair efficiency¹⁴, would also alleviate the sensitivities of RAD6 pathway mutants. Figure 1b, f shows that this was indeed the case.

The Pol η - and Pol ζ -dependent pathways of translesion synthesis operate independently of the RAD5-mediated error-free branch^{7,9,15}, resulting in additive or synergistic relationships between mutations that inactivate the individual systems. Consequently, the difference in sensitivities between *rad18 siz1* and *rad5 siz1* mutants (Fig. 1b, f) should be due to the two translesion synthesis pathways. Whereas Pol η carries out error-free replication through different types of oxidative or ultraviolet-induced lesions¹⁶ and is known to require direct interaction with PCNA for activity¹⁷, Pol ζ is most efficient in the error-prone extension of primer termini opposite a variety of lesions or mismatches and is responsible for virtually all damage-induced mutagenesis¹⁸. We asked whether Rad6p and Rad18p control translesion synthesis by means of ubiquitination of PCNA or by modification of a different target protein. We therefore analysed the effect of *rad30* and *rev3* deletions on the DNA repair efficiencies of the PCNA mutants. Deletion of *rad30* had no further effect on the ultraviolet sensitivity of the *pol30(K164R)* mutant (Fig. 1c), implying that the effect of Pol η on repair depends on the presence of K164 of PCNA. The same epistatic relationship was observed in the absence of SUMO modification in *pol30(K127/164R)* and in the *siz1* background (Fig. 1c). Thus, we conclude that it is indeed K164 of PCNA and not some other target protein whose ubiquitination is required for the repair activity of Pol η . From the fact that translesion synthesis is independent of the RAD5-mediated branch^{7,9,15}, we infer that the modification required for Pol η activity is mono- and not multi-ubiquitination. Analysis of the relationship between PCNA and Pol ζ yielded virtually identical results (Fig. 1d, g). Following the same logic as in the case of Pol η , we conclude that Pol ζ -dependent repair activity also requires mono-ubiquitination of PCNA.

If translesion synthesis by Pol η and Pol ζ fully accounts for the contribution of mono-ubiquitinated PCNA to DNA repair, deletion of both polymerases in combination with a defect in multi-ubiquitination should result in a strain with the same damage sensitivity as a PCNA mutant that is completely devoid of ubiquitination. In order to exclude the effects of SUMO we tested this notion in a *siz1* background. Multi-ubiquitination was prevented by using the *ubi(K63R)* mutant, whose repair defect equals that of a *ubc13* or *mms2* deletion^{8,19}. Figure 1e shows that indeed the mutant *ubi(K63R) rad30 rev3 siz1* had the same ultraviolet sensitivity as *pol30(K164R) siz1*, and no further increase was observed in combinations of *pol30(K164R)* with the other mutations. Treatment with DNA-damaging chemicals yielded identical results (not shown). We conclude that the phenotype of the PCNA lysine mutant is caused by a combination of defects that result on one hand from the abrogation of the error-free bypass system, dependent on K63-linked multi-ubiquitin chains, and on the other hand from inhibition of the two translesion polymerases Pol η and Pol ζ by the absence of PCNA mono-ubiquitination. On the basis of our findings, it is likely that these three systems, Pol η , Pol ζ and the multi-ubiquitin-dependent branch, account for most, if not all of the effects of PCNA ubiquitination on DNA repair.

As Pol ζ -mediated translesion synthesis is the major cause of

damage-induced mutagenesis¹⁸, we asked whether mutants deficient in the enzymes involved in ubiquitin or SUMO conjugation or whether mutations affecting the acceptor lysines on PCNA would have an impact on the accumulation of ultraviolet- and methyl methane sulphonate (MMS)-induced mutations (Fig. 2). Whereas *siz1* exhibited no defect in damage-induced mutagenesis either alone or in combination with *ubc13*, mutation of K164 of PCNA completely inhibited induced mutagenesis, similar to the *rev3* mutant itself. On the basis of the notion that *ubc13* and *siz1* are capable of mutation induction, whereas *pol30(K164R)*—similar to *rad6* and *rad18* mutants^{20,21}—is not, we conclude that mono-ubiquitination of PCNA is a prerequisite for Pol ζ -dependent damage-induced mutagenesis.

Pol ζ is also responsible for 50–75% of all spontaneous mutations¹⁸. In light of this fact the defect of the *rad6* and *rad18* mutants in many assays of Pol ζ -dependent induced mutagenesis is hard to reconcile with the notion that both display a spontaneous Pol ζ -dependent hypermutator phenotype^{22–24}. In order to examine this phenomenon, we determined spontaneous mutation rates in two independent reporter genes (Table 1). As noted before^{22,24,25}, *rad18* and *ubc13* mutants both exhibited elevated mutation rates, whereas *rev3* had a three- to fivefold reduced level of mutagenesis. Notably, *pol30(K164R)* behaved like the wild type, indicating that in contrast to induced mutagenesis, wild-type levels of spontaneous mutagenesis do not require ubiquitination of PCNA. Interestingly, however, *pol30(K164R)* completely abolished the hypermutator

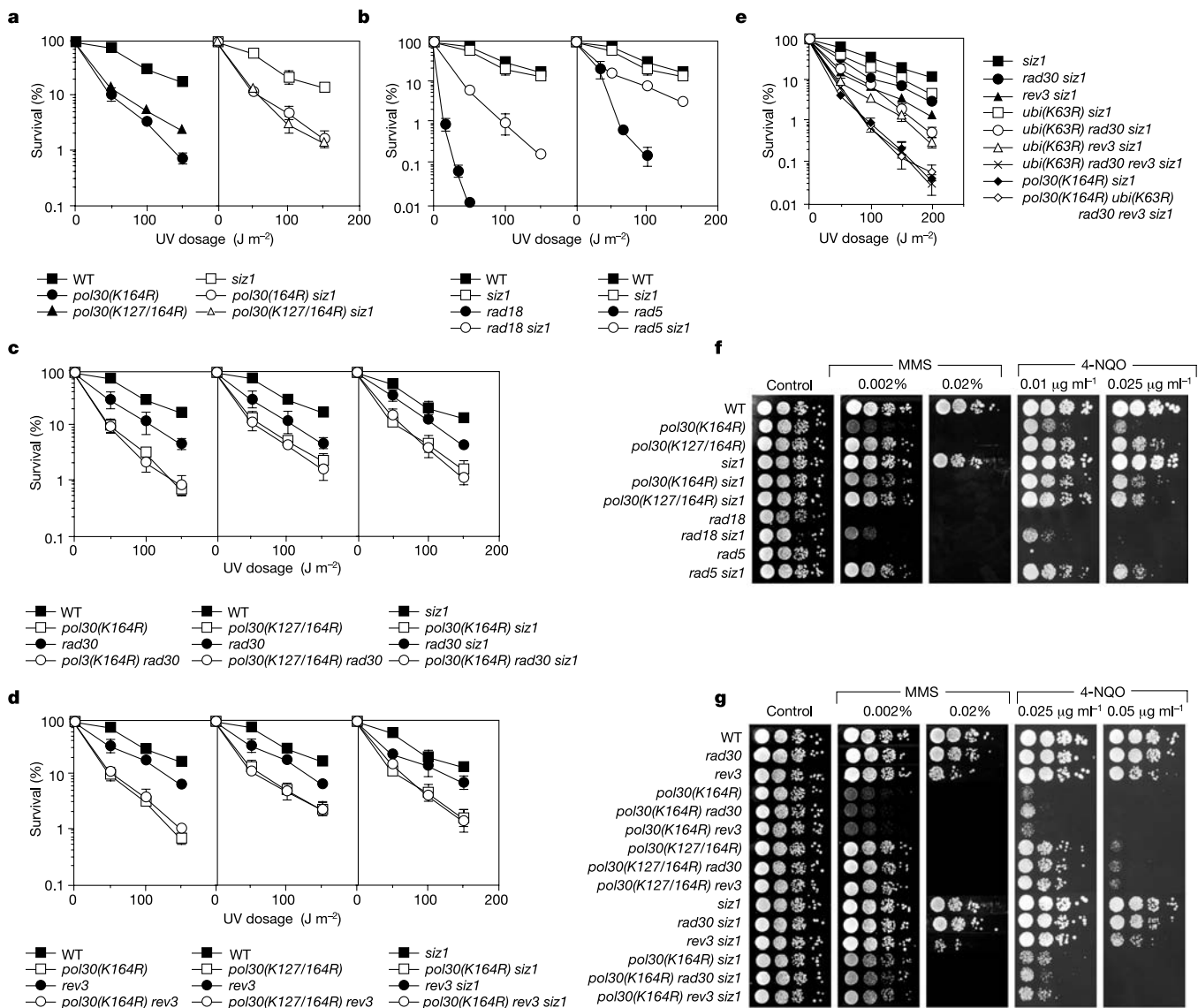


Figure 1 PCNA modifications differentially affect individual branches of the *RAD6* pathway. **a**, Effect of *siz1* on the ultraviolet (UV) sensitivities of PCNA mutants. **b**, Effect of *siz1* on the ultraviolet sensitivities of *RAD6* pathway mutants. **c**, Effect of PCNA mutants on Pol η -dependent repair. **d**, Effect of PCNA mutants on Pol ζ -dependent repair. **e**, Comparison of the ultraviolet sensitivities of strains deficient in individual branches of the *RAD6* pathway and combinations thereof. **f**, Sensitivities to MMS and 4-nitroquinoline-1-oxide (4-NQO) of the same set of strains as in **a** and **b**. **g**, Sensitivities to MMS and 4-NQO of the same set of strains as in **c** and **d**. Mutants, created as described previously⁹, were derived from a strain in which a *pol30* deletion was complemented by wild type or mutant *POL30* under the control of its own promoter on an integrative plasmid, except for

the series of strains used in **e**, which were created in a strain background where endogenous ubiquitin genes were replaced by wild-type ubiquitin or *ubi(K63R)* on a plasmid¹⁹. Here the *pol30(K164R)* allele was introduced by direct replacement of *POL30*. Details of strain construction are available on request. Survival after ultraviolet irradiation (254 nm) was determined in quadruplicate as previously described⁹. Standard deviations are indicated where the error bars exceed the sizes of the plot symbols. Sensitivities to DNA-damaging chemicals were determined by spotting tenfold serial dilutions of exponential yeast cultures of equal densities onto plates of rich medium containing the indicated concentrations of the relevant agent. Pictures were taken after incubation for 2–3 days.

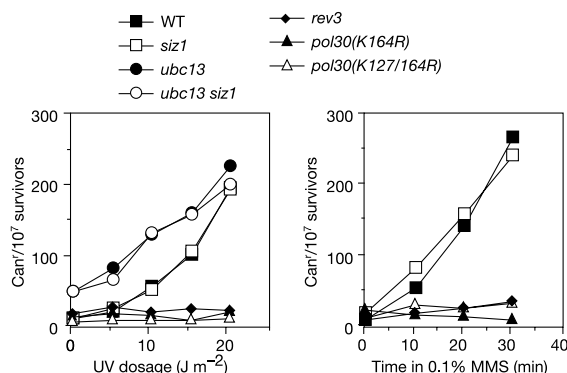


Figure 2 Pol ζ -dependent DNA-damage-induced mutagenesis requires mono-ubiquitination of PCNA. Forward mutation frequencies in the *CAN1* locus²⁰ were determined in duplicate after irradiation of the indicated strains with increasing ultraviolet dosage at 254 nm (left panel) or treatment with 0.1% MMS in liquid culture (right panel) by relating the number of colonies on minimal medium containing 30 mg l⁻¹ canavanine sulphate (Can^r) to the total number of survivors on non-selective medium. The *ubc13* derivatives⁹ exhibit a higher frequency of mutants even in the absence of DNA damage due to an elevated spontaneous mutation rate (Table 1). Both PCNA mutants in combination with any other *RAD6* pathway mutant were also completely defective in ultraviolet-induced mutagenesis (not shown).

phenotype of both *rad18* and *ubc13*, reducing it down to the level of the wild type or even below that. Thus, modification of this residue apparently is responsible for the elevated mutation rates observed in *rad18* and *ubc13* mutants. Moreover, the double-lysine mutant *pol30(K127/164R)* consistently yielded even lower mutation rates than the wild type either in isolation or in combination with *rad18* or *ubc13*. This implies that SUMO modification of PCNA contributes to spontaneous mutagenesis. We predicted that in this case deletion of *SIZ1* should negatively affect spontaneous mutation rates. In fact, although the *siz1* deletion in isolation did not affect mutagenesis, it reduced the mutation rate of *rad18* cells down to the level of *pol30(K127/164R)*. In *ubc13* mutants the *siz1* deletion reduced the elevated spontaneous mutation rate only partially. We conclude that Pol ζ -dependent spontaneous mutagenesis can be stimulated by either mono-ubiquitination or SUMO modification of PCNA.

Reduced mutation rates are seen only when neither SUMO nor ubiquitin can be attached to PCNA (as in *pol30(K127/164R)* and in *rad18 siz1*). Modification by either ubiquitin at K164 or SUMO at K164 or K127 is sufficient to afford wild-type levels of mutagenesis

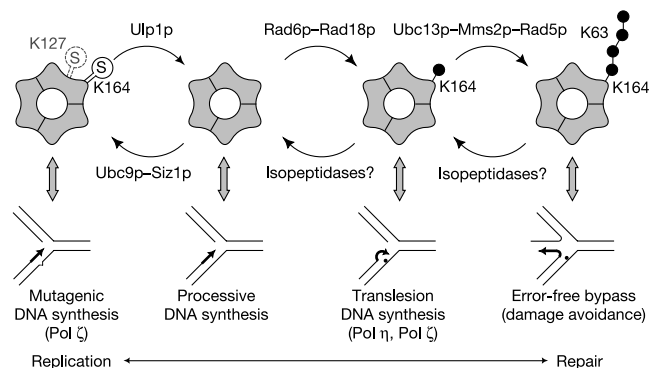


Figure 3 Model of the interplay between the SUMO and ubiquitin conjugation systems and their consequences for DNA replication and repair. PCNA is modified alternatively by ubiquitin or SUMO as depicted by black and white lollipop symbols, respectively. Lysine residues on PCNA and ubiquitin used for conjugation are indicated. Whereas K63-linked multi-ubiquitin chains synthesized by Ubc13p, Mms2p and Rad5p promote error-free bypass by a damage avoidance mechanism possibly involving a template switch to the newly synthesized sister chromatid (fourth column), mono-ubiquitination of PCNA by Rad6p and Rad18p is essential for Pol η - and Pol ζ -dependent translesion DNA synthesis, resulting in the accumulation of damage-induced mutations (third column, see also Figs 1 and 2). During normal DNA replication Pol ζ is stimulated by SUMO modification of PCNA, generating spontaneous mutations in the absence of DNA damage (Table 1, first column); in this mode Pol ζ activity is not damage-inducible. In the absence of exogenous damage, mono-ubiquitin can partially substitute for SUMO in the activation of Pol ζ (Table 1). We postulate a reversible distribution between the various modification states of PCNA, which would require the action of the relevant isopeptidases in addition to the conjugation machinery (see Supplementary Fig. 2).

(as in *pol30(K164R)* and in *siz1*), indicating that SUMO and ubiquitin can partially substitute for each other. Residual Pol ζ activity in the presence of unmodified PCNA may explain the difference remaining between *rev3* and *pol30(K127/164R)*. Whereas SUMO, mono-ubiquitin and multi-ubiquitin chains normally compete for modification of PCNA, elevated spontaneous mutation rates seem to result from an imbalance in this distribution: in *rad18*, SUMO modification predominates in the absence of ubiquitination; consequently, the elevated mutation rate is completely abolished by abrogation of SUMO modification on deletion of *SIZ1*. In contrast, the hypermutator phenotype of *ubc13* is of twofold origin: as this mutant is incapable of multi-ubiquitination, mono-ubiquitin and SUMO are the only possible PCNA modifications, both contributing to mutagenesis; here only the SUMO-dependent portion is abolished in the *siz1* background. Consistent

Table 1 Ubiquitination and SUMO modification of PCNA contribute to spontaneous mutagenesis

Strain	CAN1 forward assay		his1-1 reversion assay	
	Mutation rate $\times 10^7$	Relative rate	Mutation rate $\times 10^9$	Relative rate
Wild type	3.1 $\pm$ 0.8	1.0	1.4 $\pm$ 0.5	1.0
<i>rad18</i>	14.4 $\pm$ 4.3	4.6	3.9 $\pm$ 1.2	2.8
<i>ubc13</i>	12.9 $\pm$ 3.8	4.1	4.5 $\pm$ 1.3	3.2
<i>rev3</i>	0.9 $\pm$ 0.2	0.3	0.3 $\pm$ 0.2*	0.2
<i>pol30(K164R)</i>	3.4 $\pm$ 0.9	1.1	1.2 $\pm$ 0.4*	0.9
<i>pol30(K164R) rad18</i>	3.3 $\pm$ 0.8	1.1	0.6 $\pm$ 0.2*	0.4
<i>pol30(K164R) ubc13</i>	2.9 $\pm$ 0.7	0.9	0.8 $\pm$ 0.3*	0.6
<i>pol30(K127/164R)</i>	1.9 $\pm$ 0.5	0.6	0.7 $\pm$ 0.2*	0.5
<i>pol30(K127/164R) rad18</i>	2.0 $\pm$ 0.5	0.6	0.3 $\pm$ 0.1*	0.2
<i>pol30(K127/164R) ubc13</i>	2.0 $\pm$ 0.5	0.6	0.7 $\pm$ 0.3*	0.5
<i>siz1</i>	3.1 $\pm$ 0.8	1.0	1.4 $\pm$ 0.5	1.0
<i>rad18 siz1</i>	2.3 $\pm$ 0.6	0.7	0.7 $\pm$ 0.2*	0.5
<i>ubc13 siz1</i>	5.2 $\pm$ 1.4	1.7	2.3 $\pm$ 0.7	1.6

Spontaneous mutation rates per generation were determined in the *CAN1* and *his1-1* locus by a modified fluctuation analysis²⁹. The *can1* mutants were detected as described in the legend to Fig. 2. Revertants of *his1-1* were detected on histidine-free medium. For this purpose, the *his1-1* ochre allele, which is not subject to suppression by extragenic mutations³⁰, was introduced by replacement of *HIS1*. Absolute rates were calculated from a total of 11 (*CAN1*) or 22 (*his1-1*) cultures per strain by the method of the median except where indicated. Relative rates are given with respect to the wild type.

* Absolute rates were calculated based on the fraction of cultures without mutants²⁹.

with the stimulating effect of SUMO on spontaneous mutagenesis, we found a slightly increased rate (1.3-fold) in the *ulp1^{ts}* isopeptidase mutant^{26,27}, which displays a partial defect in SUMO deconjugation of PCNA at the permissive temperature (see Supplementary Fig. 2).

Our findings lead us to propose a model for the interplay between ubiquitin and SUMO conjugation during DNA replication and repair (Fig. 3): according to ref. 6, PCNA acts as a molecular switch that in its SUMO-modified form may promote replication, whereas multi-ubiquitination stimulates error-free repair. We have now shown that mono-ubiquitination of PCNA is essential for translesion DNA synthesis by Pol η and Pol ζ , resulting in the accumulation of damage-induced mutations. In addition, Pol ζ is also involved in spontaneous mutagenesis during DNA replication. For this purpose Pol ζ can be stimulated by SUMO modification of PCNA; however, in this replicative mode its activity is not inducible by DNA damage, thus explaining the puzzling mutagenesis phenotypes of *rad6* and *rad18* strains. Our model involving differential activation of Pol ζ by ubiquitinated and SUMO-modified PCNA for mutagenesis is consistent with a twofold origin of mutations: according to current beliefs, spontaneous mutations arise not only from replication across unrepaired lesions²², but also from Pol ζ -dependent extension of terminal mismatches, hairpins or other structural features of template DNA difficult to overcome by purely replicative polymerases¹⁸. Thus, we postulate that one function of SUMO during normal S phase is to harness Pol ζ to overcome replication fork blocks not caused by damage but by other refractory DNA structures. Activation of the repair polymerases by PCNA could be explained by a preferential interaction with the modified forms of PCNA. However, as unmodified, recombinant PCNA interacts productively with Pol η *in vitro*¹⁷ and we find that even the PCNA lysine mutants are not impaired in their interaction with Pol η (not shown), we consider it more likely that ubiquitin and SUMO may be responsible for the dissociation of other PCNA-binding proteins, such as replicative polymerases²⁸, to allow access of the translesion polymerases to the primer terminus. Alternatively, the modifications might not affect binding of the polymerases at all, but could exert a more subtle, modulating effect on their activity or processivity. Future studies will have to address the molecular mechanisms that regulate the balance between SUMO modification, mono-ubiquitination and multi-ubiquitination of PCNA and thereby control the accuracy of DNA replication and repair. □

Received 19 June; accepted 24 July 2003; doi:10.1038/nature01965.

- Herskovic, A. & Ciechanover, A. The ubiquitin system. *Annu. Rev. Biochem.* **67**, 425–479 (1998).
- Hicke, L. Gettin' down with ubiquitin: turning off cell-surface receptors, transporters and channels. *Trends Cell Biol.* **9**, 107–112 (1999).
- Conaway, R. C., Brower, C. S. & Conaway, J. W. Emerging roles of ubiquitin in transcription regulation. *Science* **296**, 1254–1258 (2002).
- Ulrich, H. D. Degradation or maintenance: actions of the ubiquitin system on eukaryotic chromatin. *Eukaryot. Cell* **1**, 1–10 (2002).
- Melchior, F. SUMO—nonclassical ubiquitin. *Annu. Rev. Cell Dev. Biol.* **16**, 591–626 (2000).
- Hoegge, C., Pfander, B., Moldovan, G. L., Pyrowolakis, G. & Jentsch, S. RAD6-dependent DNA repair is linked to modification of PCNA by ubiquitin and SUMO. *Nature* **419**, 135–141 (2002).
- Broomfield, S., Chow, B. L. & Xiao, W. MMS2, encoding a ubiquitin-conjugating-enzyme-like protein, is a member of the yeast error-free postreplication repair pathway. *Proc. Natl Acad. Sci. USA* **95**, 5678–5683 (1998).
- Hofmann, R. M. & Pickart, C. M. Noncanonical MMS2-encoded ubiquitin-conjugating enzyme functions in assembly of novel polyubiquitin chains for DNA repair. *Cell* **96**, 645–653 (1999).
- Ulrich, H. D. & Jentsch, S. Two RING finger proteins mediate cooperation between ubiquitin-conjugating enzymes in DNA repair. *EMBO J.* **19**, 3388–3397 (2000).
- Lawrence, C. W. The RAD6 DNA repair pathway in *Saccharomyces cerevisiae*: What does it do, and how does it do it? *Bioessays* **16**, 253–258 (1994).
- Bailly, V., Lauder, S., Prakash, S. & Prakash, L. Yeast DNA repair proteins Rad6 and Rad18 form a heterodimer that has ubiquitin conjugating, DNA binding, and ATP hydrolytic activities. *J. Biol. Chem.* **272**, 23360–23365 (1997).
- Jentsch, S., McGrath, J. P. & Varshavsky, A. The yeast DNA repair gene RAD6 encodes a ubiquitin-conjugating enzyme. *Nature* **329**, 131–134 (1987).
- Sung, P., Prakash, S. & Prakash, L. Mutation of cysteine-88 in the *Saccharomyces cerevisiae* RAD6 protein abolishes its ubiquitin-conjugating activity and its various biological functions. *Proc. Natl Acad. Sci. USA* **87**, 2695–2699 (1990).
- Johnson, E. & Gupta, A. A. An E3-like factor that promotes SUMO conjugation to the yeast septins. *Cell* **106**, 735–744 (2001).
- McDonald, J. P., Levine, A. S. & Woodgate, R. The *Saccharomyces cerevisiae* RAD30 gene, a homologue

- of *Escherichia coli* *dinB* and *umuC*, is DNA damage inducible and functions in a novel error-free postreplication repair mechanism. *Genetics* **147**, 1557–1568 (1997).
- Prakash, S. & Prakash, L. Translesion DNA synthesis in eukaryotes: A one- or two-polymerase affair. *Genes Dev.* **16**, 1872–1883 (2002).
- Haracska, L., Kondratieva, C. M., Unk, I., Prakash, S. & Prakash, L. Interaction with PCNA is essential for yeast DNA Polymerase η function. *Mol. Cell* **8**, 407–415 (2001).
- Lawrence, C. W. & Maher, V. M. Mutagenesis in eukaryotes dependent on DNA polymerase zeta and Rev1p. *Phil. Trans. R. Soc. Lond. B* **356**, 41–46 (2001).
- Spence, J., Sadi, S., Haas, A. L. & Finley, D. A ubiquitin mutant with specific defects in DNA repair and multiubiquitination. *Mol. Cell. Biol.* **15**, 1265–1273 (1995).
- Cassier-Chauvat, C. & Fabre, F. A similar defect in UV-induced mutagenesis conferred by the *rad6* and *rad18* mutations of *Saccharomyces cerevisiae*. *Mutat. Res.* **254**, 247–253 (1991).
- Lawrence, C. W. & Christensen, R. B. UV-mutagenesis in radiation-sensitive strains of yeast. *Genetics* **82**, 207–232 (1976).
- Hastings, P. J., Quah, S.-K. & Von Borstel, R. C. Spontaneous mutation by mutagenic repair of spontaneous lesions. *Nature* **264**, 719–722 (1976).
- Morrison, A. *et al.* REV3, a *Saccharomyces cerevisiae* gene whose function is required for induced mutagenesis, is predicted to encode a non-essential DNA polymerase. *J. Bacteriol.* **171**, 5659–5667 (1989).
- Quah, S.-K., Von Borstel, R. C. & Hastings, P. J. The origin of spontaneous mutation in *Saccharomyces cerevisiae*. *Genetics* **96**, 819–839 (1980).
- Brusky, J., Zhu, Y. & Xiao, W. UBC13, a DNA-damage-inducible gene, is a member of the error-free postreplication repair pathway in *Saccharomyces cerevisiae*. *Curr. Genet.* **37**, 168–174 (2000).
- Li, S.-J. & Hochstrasser, M. A new protease required for cell-cycle progression in yeast. *Nature* **398**, 246–251 (1999).
- Li, S.-J. & Hochstrasser, M. The yeast ULP2 (*SMT4*) gene encodes a novel protease specific for the ubiquitin-like Smt3 protein. *Mol. Cell. Biol.* **20**, 2367–2377 (2000).
- Warbrick, E. The puzzle of PCNA's many partners. *Bioessays* **22**, 997–1006 (2000).
- Lea, D. E. & Coulson, C. A. The distribution of the numbers of mutants in bacterial populations. *J. Genet.* **49**, 264–285 (1949).
- Von Borstel, R. C. *et al.* Topical reversion at the *HIS1* locus of *Saccharomyces cerevisiae*: a tale of three mutants. *Genetics* **148**, 1647–1654 (1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M. Ludwig for technical assistance; R. Viana for construction of the His₆-POL30 vector used for the Supplementary Information; M. Hochstrasser for the *ulp1^{ts}* and *ulp2* mutants; C. Hoegge and S. Jentsch for the anti-PCNA antibody used for the Supplementary Information; and W. Kramer for comments on the manuscript. R. Kahmann is acknowledged for generous support. H.D.U. is supported by grants from the German Ministry for Education and Research, Deutsche Forschungsgemeinschaft and the German-Israeli Foundation for Scientific Research and Development.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.D.U. (ulrich@staff.uni-marburg.de).

Small molecule activators of sirtuins extend *Saccharomyces cerevisiae* lifespan

Konrad T. Howitz¹, Kevin J. Bitterman², Haim Y. Cohen², Dudley W. Lamming², Siva Lavu², Jason G. Wood², Robert E. Zipkin¹, Phuong Chung¹, Anne Kisieleski¹, Li-Li Zhang¹, Brandy Scherer¹ & David A. Sinclair²

¹BIOMOL Research Laboratories, Inc., 5120 Butler Pike, Plymouth Meeting, Pennsylvania 19462, USA

²Department of Pathology, Harvard Medical School, 200 Longwood Avenue, Boston, Massachusetts 02115, USA

In diverse organisms, calorie restriction slows the pace of ageing and increases maximum lifespan. In the budding yeast *Saccharomyces cerevisiae*, calorie restriction extends lifespan by increasing the activity of Sir2 (ref. 1), a member of the conserved sirtuin family of NAD⁺-dependent protein deacetylases^{2–6}. Included in this family are SIR2.1, a *Caenorhabditis elegans* enzyme that regulates lifespan⁷, and SIRT1, a human deacetylase that promotes cell survival by negatively regulating the p53 tumour

suppressor^{8–10}. Here we report the discovery of three classes of small molecules that activate sirtuins. We show that the potent activator resveratrol, a polyphenol found in red wine, lowers the Michaelis constant of SIRT1 for both the acetylated substrate and NAD⁺, and increases cell survival by stimulating SIRT1-dependent deacetylation of p53. In yeast, resveratrol mimics calorie restriction by stimulating Sir2, increasing DNA stability and extending lifespan by 70%. We discuss possible evolutionary origins of this phenomenon and suggest new lines of research into the therapeutic use of sirtuin activators.

Longevity regulatory genes have been identified in many eukaryotes, including rodents, flies, nematode worms and even single-celled organisms such as baker's yeast (reviewed in ref. 11). These genes seem to be part of an evolutionarily conserved longevity pathway that evolved to promote survival in response to deteriorating environmental conditions. We recently reported that the pace of ageing in yeast is governed by *PNC1*, a calorie restriction-responsive and stress-responsive gene that depletes nicotinamide¹², a physiological inhibitor of Sir2¹³. On the basis of these results we proposed that calorie restriction might confer health benefits because it is a mild stress that provokes an organismal defence response¹².

On the basis of the positive health effects of calorie restriction in mammals¹⁴, we sought to identify small molecules that could modulate sirtuin activity. To this end, we screened a number of small molecule libraries, which included analogues of ϵ -acetyl lysine, NAD⁺, NAD⁺ precursors, nucleotides and purinergic ligands. Fluorescent deacetylation assays were performed on human SIRT1 using a synthetic peptide substrate that encompassed lysine 382 (K382) of p53, a known target of SIRT1 *in vivo*^{8–10} and a preferred target *in vitro* (see Supplementary Fig. 1).

The initial compound screen identified several inhibitors of SIRT1 (Supplementary Table 7). However, the most notable outcome was the identification of two structurally similar compounds, quercetin and piceatannol, that stimulated SIRT1 activity five- and eightfold, respectively (Table 1). Both quercetin and piceatannol have previously been identified as protein kinase inhibitors^{15,16}. Comparison of the two compounds suggested a possible structure-activity relationship. The *trans*-stilbene ring structures of piceatannol are superimposable on the flavonoid A and B rings of quercetin, with the ether oxygen and carbon-2 of the C ring aligning with the ethylene carbons in piceatannol (Table 1). Furthermore, the 5, 7, 3' and 4' hydroxyl group positions in quercetin can be aligned, respectively, with the 3, 5, 3' and 4' hydroxyls of piceatannol.

Both quercetin and piceatannol are polyphenols, members of a large and diverse group of plant secondary metabolites that include flavones, stilbenes, flavanones, isoflavones, catechins (flavan-3-ols), chalcones, tannins and anthocyanidins^{17,18}. A secondary screen of these compound families identified 15 additional SIRT1 activators that provided more than twofold stimulation (Table 1; see also Supplementary Tables 1, 2 and 4). We refer to the sirtuin-activating compounds as STACs.

Many, but not all of the most active STACs included hydroxyls in the two meta-positions of the A ring, *trans* to the B ring with a 4' or 3',4' hydroxyl pattern (Table 1; see also Supplementary Tables 1 and 2). A potentially coplanar orientation and *trans*-positioning of the hydroxylated rings also appear to be important for activity. The most potent activator was resveratrol (3,5,4'-trihydroxystilbene), a polyphenol found in red wine that is associated with a surprising number of health benefits, most notably the mitigation of age-related diseases, including neurodegeneration, carcinogenesis and atherosclerosis^{17–19}.

The biological effects of polyphenols are frequently attributed to antioxidant, metal-ion-chelating and/or free-radical-scavenging activity¹⁷. We considered the possibility that the stimulation of SIRT1 might simply represent the repair of oxidative- or metal-ion-induced damage to the recombinant protein. We discounted this

explanation based on three observations. First, a variety of free-radical protective compounds, including antioxidants, chelators and radical scavengers failed to stimulate SIRT1 (see Supplementary Table 6). Second, among various polyphenols of equivalent antioxidant capacity²⁰ we observed diverse SIRT1-stimulating activity. Third, as described below, we showed that resveratrol could stimulate the activity of native sirtuins *in vivo*.

Dose-response experiments showed that resveratrol doubled the rate of deacetylation by SIRT1 at about 11 μ M and was saturated at 100–200 μ M (Fig. 1a). Although resveratrol had no significant effect on the two $V_{\max}$ determinations when either substrate or NAD⁺ was varied (Fig. 1b, c), it had pronounced effects on the two apparent Michaelis constants (K_m). Its effect on the acetylated peptide K_m was particularly striking, amounting to a 35-fold

Table 1 Stimulation of SIRT1 catalytic rate by plant polyphenols

Compound (100 μ M)	Ratio to control (mean $\pm$ s.e.)	Structure
Resveratrol (3,5,4'-trihydroxy- <i>trans</i> -stilbene)	13.4 $\pm$ 1.0	
Butein (3,4,2',4'-tetrahydroxychalcone)	8.53 $\pm$ 0.89	
Piceatannol (3,5,3',4'-tetrahydroxy- <i>trans</i> -stilbene)	7.90 $\pm$ 0.50	
Isoliquiritigenin (4,2',4'-trihydroxychalcone)	7.57 $\pm$ 0.84	
Fisetin (3,7,3',4'-tetrahydroxyflavone)	6.58 $\pm$ 0.69	
Quercetin (3,5,7,3',4'-pentahydroxyflavone)	4.59 $\pm$ 0.47	

Rate measurements with 25 μ M NAD⁺ and 25 μ M p53-382 acetylated peptide substrate were performed as described in Methods. All ratio data were calculated from experiments in which the total deacetylation in the control reaction was 0.25–1.25 μ M peptide or 1–5% of the initial concentration of acetylated peptide. s.e., standard error.

decrease (Fig. 1b). Resveratrol also lowered the K_m for NAD^+ over fivefold (Fig. 1c). As resveratrol acts only on K_m , it could be classified as an allosteric effector of the 'K system' type²¹, which may indicate that only the substrate-binding affinity of the enzyme has been altered. In the presence of the negative regulator nicotinamide, resveratrol stimulated SIRT1, but there also appeared to be complex, concentration-dependent effects (Fig. 1d). Protein struc-

tural studies are underway to determine how resveratrol activates SIRT1 and to elucidate the interaction between the effects of nicotinamide and polyphenols.

To investigate the effects of sirtuin activators *in vivo*, we turned to *S. cerevisiae*. In this organism, a single extra copy of *SIR2* is sufficient to mimic calorie restriction¹. We determined that the optimal resveratrol concentration for activation of recombinant Sir2 was 2–5 μM and that, in contrast to recombinant SIRT1, the rate stimulation was no more than about twofold (Fig. 2a). Resveratrol and other STACs in the stilbene, flavone and chalcone families were then tested for their effect on yeast replicative lifespan²². Three compounds—butein, fisetin and resveratrol—increased average lifespan by 31%, 55% and 70%, respectively, and all three significantly increased maximum lifespan (Fig. 2b). Higher concentrations of resveratrol provided no added lifespan benefit (Fig. 2c), and there was no lasting effect of the compound on the lifespan of pre-treated young cells (data not shown). Quercetin and piceatannol had no

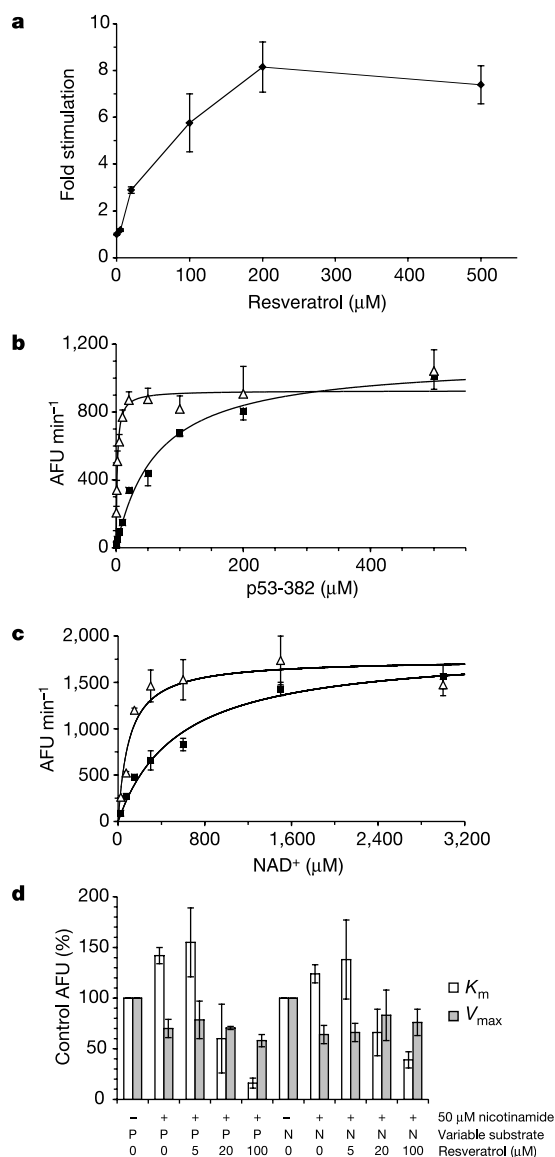


Figure 1 Effects of resveratrol on the kinetics of recombinant SIRT1. **a**, Resveratrol dose-response of SIRT1 catalytic rate. Fold stimulation values are derived from the means of six to nine determinations. **b**, SIRT1 initial rate at 3 mM NAD^+ , as a function of p53-382 acetylated peptide concentration in the presence (triangles) or absence (squares) of 100 μM resveratrol. Lines represent nonlinear least-square fits to the Michaelis-Menten equation. K_m (control, squares) = 64 μM , K_m (plus resveratrol, triangles) = 1.8 μM ; V_{max} (control, squares) = 1,107 AFU min⁻¹, V_{max} (plus resveratrol, triangles) = 926 AFU min⁻¹. **c**, SIRT1 initial rate at 1 mM p53-382 acetylated peptide, as a function of NAD^+ concentration, in the presence (triangles) or absence (squares) of 100 μM resveratrol as in **b**. K_m (control, squares) = 558 μM , K_m (plus resveratrol, triangles) = 101 μM ; V_{max} (control, squares) = 1,863 AFU min⁻¹, V_{max} (plus resveratrol, triangles) = 1,749 AFU min⁻¹. **d**, Kinetic constants are relative to untreated control and represent the mean of two determinations. AFU, arbitrary fluorescence unit; N, NAD^+ ; P, p53 acetylated peptide.

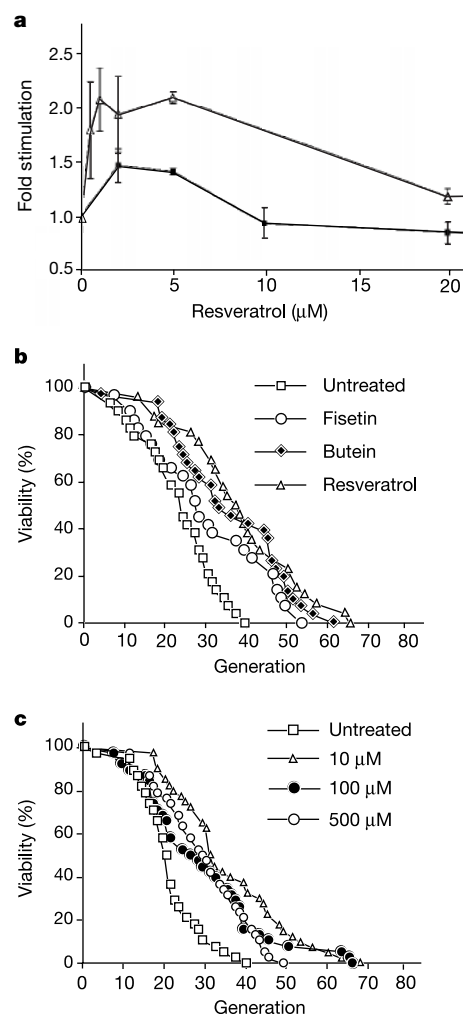


Figure 2 Effects of polyphenols on Sir2 and *S. cerevisiae* lifespan. **a**, Rates were determined using either FdL acetylated lysine substrate (squares) or an acetylated fluorogenic peptide (triangles) based on the sequence encompassing histone H4 Lys 16, a target of Sir2 *in vivo* (Supplementary Fig. 1). Fold stimulations represent the mean of at least three determinations. **b**, Replicative lifespan is the number of daughter cells an individual mother cell produces before dying. Lifespan analyses were determined by micro-manipulating individual yeast cells²³ on complete 2% glucose medium with 10 μM compound, unless otherwise stated. Average lifespan for wild type, 22.9 generations; fisetin, 30.0; butein, 35.5; resveratrol, 36.8. **c**, Average lifespan for wild type untreated, 21.0 generations; grown on resveratrol 10 μM , 35.7; 100 μM , 29.4; 500 μM , 29.3.

significant effect on lifespan (data not shown), possibly due to oxidation of the compounds in the medium or insufficient uptake into cells.

For subsequent experiments we focused on resveratrol. Glucose restriction, a form of calorie restriction in yeast, resulted in no significant extension of long-lived, resveratrol-treated cells (Fig. 3a), indicating that resveratrol probably acts through the same pathway as calorie restriction. Consistent with this, resveratrol had no effect on the lifespan of a *sir2* null mutant (Fig. 3b). Given that resveratrol is reported to have fungicidal properties at high concentrations and that mild stress can extend yeast lifespan by activating *PNC1* (ref. 12), it was plausible that resveratrol was extending lifespan by inducing *PNC1*, rather than acting on Sir2 directly. However, resveratrol extended the lifespan of a *pnc1* null mutant nearly as well as it did wild-type cells (Fig. 3b). Together these data show that resveratrol acts downstream of *PNC1* and requires *SIR2* for its effect. Although we do not discount the possibility that resveratrol might act on Sir2 indirectly, the simplest explanation of our observations is that resveratrol increases lifespan by directly stimulating Sir2 activity *in vivo*.

The ability of Sir2 to extend yeast lifespan is thought to stem from its role in stabilizing repetitive DNA^{1,22–24}. Homologous recombination between ribosomal DNA (rDNA) repeats can generate an extrachromosomal circular DNA molecule that is replicated until it reaches toxic levels in old cells²³. Consistent with this, resveratrol reduced the frequency of rDNA recombination by about 60% in a *SIR2*-dependent manner (Fig. 3c, d). In the presence of nicotinamide, recombination was also decreased by resveratrol, in agreement with the kinetic data (Fig. 3c). Interestingly, each of the STACs we tested had only minor effects on rDNA silencing compared with the 2×*SIR2* positive control (Fig. 3e, f). Together, these findings add to the body of evidence that yeast ageing is caused by genomic instability^{1,22–25}, not gene dysregulation²⁶.

Another measure of lifespan in *S. cerevisiae* is the length of time

that cells can survive in a metabolically active but nutrient-deprived state. Ageing under these conditions (that is, chronological ageing) is primarily due to oxidative damage. Resveratrol (10 μ M or 100 μ M) failed to extend chronological lifespan (not shown), indicating that the effect of resveratrol on sirtuins may be more relevant than its antioxidant activity^{17,18}.

To test whether STACs could stimulate human sirtuins *in vivo*, we developed a cell-based fluorescence deacetylase assay (Supplementary Figs 2 and 3a–c). A selection of sirtuin-stimulatory and non-stimulatory polyphenols were analysed using this *in vivo* assay and values were plotted against their fold stimulation of SIRT1 *in vitro* (Fig. 4a). Compounds with little or no activity *in vitro* clustered around the negative control (group A). Another grouping of strong *in vitro* activators is clearly distanced from the low-activity cluster in both dimensions (group B). These data indicate that certain polyphenols can activate human sirtuins *in vivo*.

Next, we assessed whether resveratrol could activate SIRT1 *in vivo*. One known target of SIRT1 is lysine 382 of p53 (K382). Deacetylation of this residue by SIRT1 decreases the activity and half-life of p53, and increases cell survival under a variety of DNA-damaging conditions^{8–10}. Treatment of cells with a low concentration (0.5 μ M) of resveratrol increased cell survival after ionizing irradiation (Fig. 4b). To test whether this was due to modulation of SIRT1 activity, we generated an antibody that specifically recognizes the acetylated form of p53-K382 (Ac-K382) (Fig. 4c and data not shown). U2OS cells treated with 0.5 μ M resveratrol showed a marked decrease (about 75%) in the level of Ac-K382 (Fig. 4d; see also Supplementary Fig. 4a). At higher concentrations of resveratrol (>50 μ M) the effect was reversed (Fig. 4d and data not shown), which may explain the dichotomy in the literature regarding the effects of resveratrol on cell viability^{17,27,28}. Treatment of HEK 293 cells with resveratrol led to a dose-dependent decrease in Ac-K382 levels that was diminished by overexpression of a dominant-negative SIRT1 allele (see Fig. 4e and Supplementary Fig. 4b).

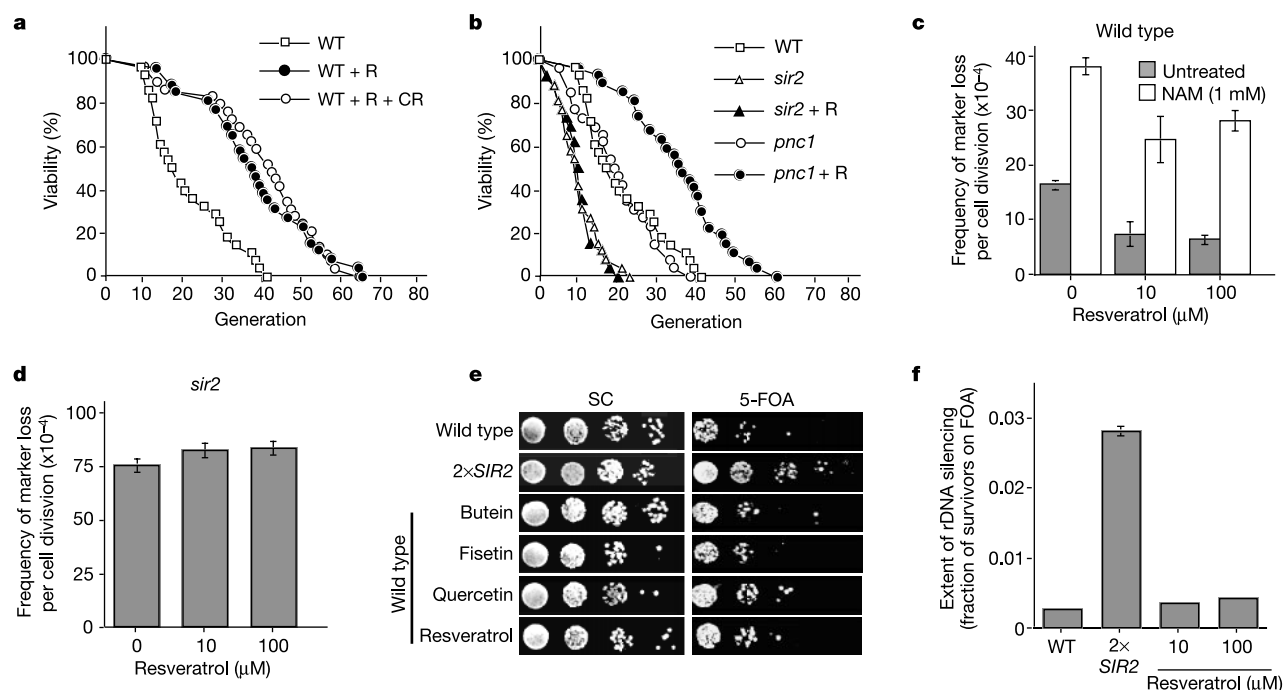


Figure 3 Resveratrol extends lifespan by mimicking calorie restriction and suppressing rDNA recombination. **a**, Average lifespan for wild type (WT) untreated, 19.0 generations; wild type plus resveratrol (WT + R), 37.8; glucose (calorie)-restricted plus resveratrol (WT + CR + R), 39.9. **b**, Average lifespan for wild type, 19.0 generations; *sir2*, 9.9; *sir2* plus resveratrol, 10.0; *pnc1*, 19.2; *pnc1* plus resveratrol, 33.1. **c**, **d**, Recombination

frequencies for wild type and *sir2* strains were determined by monitoring loss of the rDNA::ADE2 marker gene. **e**, Pre-treated *RDN1::URA3* cells were spotted as tenfold serial dilutions on SC medium with or without 5-fluoro-orotic acid (5-FOA). Increased silencing results in increased survival on 5-FOA medium. **f**, Level of rDNA silencing represents numbers of surviving cells on 5-FOA/total plated.

These data strongly indicate that resveratrol can promote cell survival by stimulating SIRT1 *in vivo*.

We have discovered small molecule sirtuin activators (STACs), and have shown that they can promote the survival of eukaryotic cells. Sirtuins have been found in diverse eukaryotes including fungi, protozoans, metazoans and plants²⁹, and probably evolved early in life's history¹¹. Plants produce a variety of polyphenols such as resveratrol in response to stresses including dehydration, nutrient deprivation, ultraviolet radiation and pathogens³⁰. Therefore it is plausible that plants synthesize these molecules in order to regulate a sirtuin-mediated stress response. This would be consistent with

the recently discovered relationship between environmental stress and Sir2 activity in yeast¹². Perhaps these compounds have stimulatory activity on sirtuins from fungi and animals because they mimic an endogenous activator. Alternatively, animal and fungal sirtuins may have retained or developed an ability to respond to plant stress metabolites because they are a useful indicator of a deteriorating environment and/or food supply. Consistent with these ideas, preliminary experiments indicate that resveratrol can also extend the lifespan of multicellular animals, including the nematode *C. elegans* and the fruitfly *Drosophila melanogaster* (J.G.W., D.A.S. and M. Tatar, unpublished results).

A broad range of human health benefits have been reported for plant polyphenols including cardioprotection, neuroprotection and cancer suppression^{17–19}. Interestingly, similar beneficial effects are observed for calorie-restricted rodents¹⁴. We postulate that many of the effects of polyphenols might be the result of a calorie-restriction-mimetic defence response mediated by sirtuins. Part of this response may involve suppressing p53 and delaying apoptosis to give cells additional time to repair damage and to prevent unnecessary cell death. Our results also suggest that survival and longevity at the cellular and organismal levels are intimately linked. The newly appreciated ability of polyphenols to promote survival and longevity by activating sirtuins indicates a new line of investigation into the effects of these and related molecules on age-related human diseases. □

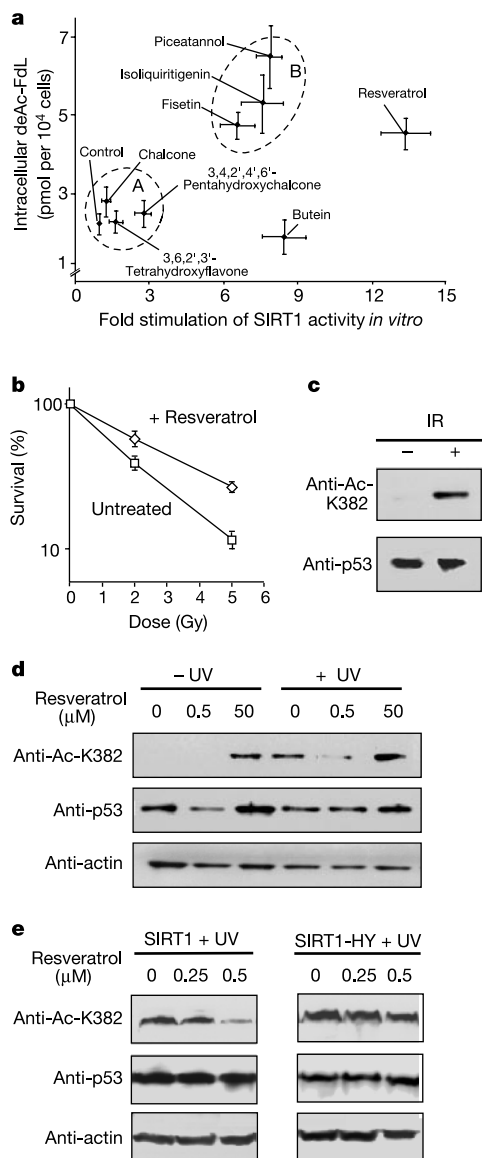


Figure 4 STACs stimulate sirtuin activity in human cells. **a**, STACs can stimulate TSA-insensitive FdL deacetylation (Supplementary Figs 2 and 3). Intracellular deacetylated FdL (deAc-FdL) levels are plotted against fold stimulation in the *in vitro* SIRT1 assay. **b**, Survival of HEK 293 cells in the presence of 0 (squares) or 0.5 μM (diamonds) resveratrol after exposure to 0, 2 or 5 Gy of ionizing radiation (IR). Cell lysates were prepared 4 h after irradiation and probed with indicated antibodies. **c**, U2OS osteosarcoma cells were exposed to 0 or 10 Gy of ionizing radiation (IR). Cell lysates were prepared 4 h after irradiation and probed with indicated antibodies. **d**, U2OS cells were pre-treated with resveratrol for 4 h after which cells were exposed to 0 or 50 J cm⁻² of ultraviolet (UV) radiation. **e**, HEK 293 cells expressing wild-type SIRT1 or dominant-negative SIRT1 H363Y (SIRT1-HY) were pre-treated with resveratrol and exposed to ultraviolet radiation as above.

Methods

Compound libraries and deacetylation assays

His₆-tagged recombinant SIRT1 and recombinant Sir2 were prepared as described¹³. A total of 0.1–1.0 μg of SIRT1 and 1.5 μg of Sir2 were used per deacetylation assay (50 μl total reaction) using 'Fluor de Lys' (FdL) as described¹³. SIRT1 screening assays used the p53-382 acetylated substrate (FdL-SIRT1, BIOMOL). Themed compound libraries (BIOMOL) were used for primary and secondary screening. Most polyphenol compounds were dissolved at 10 mM in dimethylsulphoxide (DMSO) on the day of the assay. For water-soluble compounds and negative controls, 1% v/v DMSO was added to the assay. *In vitro* fluorescence assays were read in white, half-volume 96-well microplates (Corning) with a CytoFluor II fluorescence plate reader (PerSeptive Biosystems, excitation 360 nm, emission 460 nm, gain = 85). HeLa cells were grown and the cellular deacetylation assays performed and read, as above, but in full-volume 96-well microplates (Corning). Unless otherwise indicated all initial rate measurements were means of three or more replicates, obtained with single incubation times, at which point 5% or less of the substrate initially present had been deacetylated.

For the SIRT1 resveratrol dose-response experiments, values are derived from the slopes of fluorescence (arbitrary fluorescence units, AFU) against time plots with data obtained at 0, 5, 10 and 20 min of deacetylation using 25 μM each of NAD⁺ and p53-382 acetylated peptide. Sir2 assays were performed similarly using either 100 μM FdL or 25 μM acetylated H4 peptide (FdL-H4-Ac-K16). Calculation of net fluorescence included the subtraction of a blank value by adding an inhibitor (200 μM suramin or 1 mM nicotinamide) to the reaction in the case of SIRT1, and in the case of Sir2 by omitting the enzyme from the reaction. For polyphenols that partially quenched the fluorescence produced in the assay, correction factors were obtained by determining the fluorescence increase due to a 3 μM spike of an FdL deacetylated standard (BIOMOL).

Medium and strains

All yeast strains were grown at 30 °C in complete yeast extract/bactopeptone, 2.0% (w/v) glucose (YPD) medium unless stated otherwise. Calorie restriction was induced in 0.5% glucose. Synthetic complete (SC) medium consisted of 1.67% yeast nitrogen base, 2% glucose and 40 mg l⁻¹ each of auxotrophic markers. *SIR2* was integrated in extra copy and disrupted as described¹³. Other strains are described elsewhere¹³. For cellular deacetylation assays, HeLa S3 cells were used. U2OS osteosarcoma and human embryonic kidney (HEK 293) cells were cultured adherently in DMEM containing 10% fetal calf serum with 1.0% glutamine and 1.0% penicillin/streptomycin. HEK 293 overexpressing dominant-negative SIRT1 H363Y was a gift of R. Frye.

Lifespan determinations

Lifespan measurements were performed using PSY316AT *MATα* as previously described¹². All compounds for lifespan analyses were dissolved in 95% ethanol. Plates were dried and used within 24 h. Cells were pre-incubated on their respective medium for at least 15 h and equilibrated on the medium for a minimum of 4 h before micro-manipulating them. At least 30 cells were examined per experiment and each experiment was performed at least twice. Statistical significance of lifespan differences was determined using the Wilcoxon rank sum test. Differences are stated to be significant when the confidence is higher than 95%.

Silencing and recombination assays

Ribosomal DNA silencing assays were performed as previously described¹³. Ribosomal DNA recombination frequencies were determined by plating W303AR cells²³ on YPD medium with low adenine/histidine and counting the fraction of half-red sector colonies using Bio-Rad Quantity One software. At least 6,000 cells were analysed per experiment and all experiments were performed in triplicate. All strains were pre-grown for 15 h with the relevant compound before plating.

Cellular survival assay

HEK 293 cells treated with either 0.5 μ M resveratrol or 0.5% DMSO for 8 h were trypsinized, collected and exposed to 0, 2 or 5 Gy of ionizing radiation. A total of 300 cells per sample were then plated in duplicate in 10 cm dishes. Medium was changed the following day and after 10 days, colonies were stained with crystal violet and counted. Error bars represent the standard error of the mean.

Proteins and western analyses

Recombinant Sir2–glutathione S-transferase was expressed and purified from *Escherichia coli* as previously described except that lysates were prepared using sonication¹³. Recombinant SIRT1 from *E. coli* was prepared as described¹³. Polyclonal antiserum against p53–Ac-K382 was generated using an acetylated peptide antigen as previously described⁸, with the following modifications. Anti-Ac-K382 antibody was affinity purified using non-acetylated p53–K382 peptides and stored in PBS at -70°C . Western hybridizations using anti-acetylated K382 or anti-actin (Chemicon) antibody were performed at 1:1,000 dilution of antibody. Hybridizations with polyclonal p53 antibody (Santa Cruz) used a 1:500 dilution. Cell extracts were prepared 4 h after irradiation from cells pre-treated with resveratrol for 4 h. Lysis buffer contained 150 mM NaCl, 1 mM MgCl_2 , 10% glycerol, 1% NP40, 1 mM dithiothreitol and anti-protease cocktail (Roche).

Received 13 June; accepted 31 July 2003; doi:10.1038/nature01960.

Published online 24 August 2003.

- Lin, S. J., Defossez, P. A. & Guarente, L. Requirement of NAD and SIR2 for life-span extension by calorie restriction in *Saccharomyces cerevisiae*. *Science* **289**, 2126–2128 (2000).
- Landry, J. *et al.* The silencing protein SIR2 and its homologs are NAD-dependent protein deacetylases. *Proc. Natl Acad. Sci. USA* **97**, 5807–5811 (2000).
- Imai, S., Armstrong, C. M., Kaerberlein, M. & Guarente, L. Transcriptional silencing and longevity protein Sir2 is an NAD-dependent histone deacetylase. *Nature* **403**, 795–800 (2000).
- Smith, J. S. *et al.* A phylogenetically conserved NAD⁺-dependent protein deacetylase activity in the Sir2 protein family. *Proc. Natl Acad. Sci. USA* **97**, 6658–6663 (2000).
- Tanner, K. G., Landry, J., Sternglanz, R. & Denu, J. M. Silent information regulator 2 family of NAD-dependent histone/protein deacetylases generates a unique product, 1-O-acetyl-ADP-ribose. *Proc. Natl Acad. Sci. USA* **97**, 14178–14182 (2000).
- Tanny, J. C., Dowd, G. J., Huang, J., Hilz, H. & Moazed, D. An enzymatic activity in the yeast Sir2 protein that is essential for gene silencing. *Cell* **99**, 735–745 (1999).
- Tissenbaum, H. A. & Guarente, L. Increased dosage of a sir-2 gene extends lifespan in *Caenorhabditis elegans*. *Nature* **410**, 227–230 (2001).
- Vaziri, H. *et al.* hSIRT1(SIRT1) functions as an NAD-dependent p53 deacetylase. *Cell* **107**, 149–159 (2001).
- Luo, J. *et al.* Negative control of p53 by Sir2 α promotes cell survival under stress. *Cell* **107**, 137–148 (2001).
- Langley, E. P. M. *et al.* Human SIR2 deacetylates p53 and antagonizes PML/p53-induced cellular senescence. *EMBO J.* **21**, 2383–2396 (2002).
- Kenyon, C. A conserved regulatory mechanism for ageing. *Cell* **105**, 165–168 (2001).
- Anderson, R. M., Bitterman, K. J., Wood, J. G., Medvedik, O. & Sinclair, D. A. Nicotinamide and Pnc1 govern lifespan extension by calorie restriction in *S. cerevisiae*. *Nature* **423**, 181–185 (2003).
- Bitterman, K. J., Anderson, R. M., Cohen, H. Y., Latorre-Esteves, M. & Sinclair, D. A. Inhibition of silencing and accelerated ageing by nicotinamide, a putative negative regulator of yeast sir2 and human SIRT1. *J. Biol. Chem.* **277**, 45099–45107 (2002).
- Masoro, E. J. Caloric restriction and ageing: an update. *Exp. Gerontol.* **35**, 299–305 (2000).
- Glossmann, H., Presek, P. & Eigenbrodt, E. Quercetin inhibits tyrosine phosphorylation by the cyclic nucleotide-independent, transforming protein kinase, pp60src. *Naunyn-Schmiedeberg's Arch. Pharmacol.* **317**, 100–102 (1981).
- Oliver, J. M., Burg, D. L., Wilson, B. S., McLaughlin, J. L. & Geahlen, R. L. Inhibition of mast cell Fc epsilon R1-mediated signaling and effector function by the Syk-selective inhibitor, piceatannol. *J. Biol. Chem.* **269**, 29697–29703 (1994).
- Ferguson, L. R. Role of plant polyphenols in genomic stability. *Mutat. Res.* **475**, 89–111 (2001).
- Middleton, E. Jr, Kandaswami, C. & Theoharides, T. C. The effects of plant flavonoids on mammalian cells: implications for inflammation, heart disease, and cancer. *Pharmacol. Rev.* **52**, 673–751 (2000).
- Jang, M. *et al.* Cancer chemopreventive activity of resveratrol, a natural product derived from grapes. *Science* **275**, 218–220 (1997).
- Stojanovic, S., Sprinz, H. & Brede, O. Efficiency and mechanism of the antioxidant action of trans-resveratrol and its analogues in the radical liposome oxidation. *Arch. Biochem. Biophys.* **391**, 79–89 (2001).
- Monod, J., Wyman, J. & Changeux, J.-P. On the nature of allosteric transitions. *J. Mol. Biol.* **12**, 88–118 (1965).
- Sinclair, D. A. Paradigms and pitfalls of yeast longevity research. *Mech. Ageing Dev.* **123**, 857–867 (2002).
- Sinclair, D. A. & Guarente, L. Extrachromosomal rDNA circles—a cause of ageing in yeast. *Cell* **91**, 1033–1042 (1997).
- Defossez, P. A. *et al.* Elimination of replication block protein Fob1 extends the life span of yeast mother cells. *Mol. Cell* **3**, 447–455 (1999).
- Kaerberlein, M., McVey, M. & Guarente, L. The SIR2/3/4 complex and SIR2 alone promote longevity in *Saccharomyces cerevisiae* by two different mechanisms. *Genes Dev.* **13**, 2570–2580 (1999).
- Jazwinski, S. M. Metabolic control and gene dysregulation in yeast ageing. *Ann. NY Acad. Sci.* **908**, 21–30 (2000).

- Dong, Z. Molecular mechanism of the chemopreventive effect of resveratrol. *Mutat. Res.* **523–524**, 145–150 (2003).
- Nicolini, G., Rigolio, R., Miloso, M., Bertelli, A. A. & Tredici, G. Anti-apoptotic effect of trans-resveratrol on paclitaxel-induced apoptosis in the human neuroblastoma SH-SY5Y cell line. *Neurosci. Lett.* **302**, 41–44 (2001).
- Pandey, R. *et al.* Analysis of histone acetyltransferase and histone deacetylase families of *Arabidopsis thaliana* suggests functional diversification of chromatin modification among multicellular eukaryotes. *Nucleic Acids Res.* **30**, 5036–5055 (2002).
- Soleas, G. J., Diamandis, E. P. & Goldberg, D. M. Resveratrol: a molecule whose time has come? And gone? *Clin. Biochem.* **30**, 91–113 (1997).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We wish to thank members of the Sinclair and BIOMOL laboratories for discussions and manuscript preparation, S. Luikenhuis and J. Fox for critical reading of the manuscript, and R. Frye for reagents. This work was supported by the National Institute on Aging and the Harvard-Armstrong Foundation. D.S. is an Ellison Medical Research Foundation New Research Scholar. K.B. is a Harvard Medical School Pathology Department MPM Scholar. H.C. is supported by the American Federation of Aging Research, D.L. by a National Eye Institute training grant, and J.W. by an NSF Fellowship.

Competing interests statement The authors declare competing financial interests: details accompany the paper on www.nature.com/nature.

Correspondence and requests for materials should be addressed to D.S. (david_sinclair@hms.harvard.edu).

A cell surface receptor mediates extracellular Ca^{2+} sensing in guard cells

Shengcheng Han¹, Ruhang Tang¹, Lisa K. Anderson¹, Todd E. Woerner² & Zhen-Ming Pei¹

¹Department of Biology, Developmental, Cell, and Molecular Biology Group, and

²Department of Chemistry, Duke University, Durham, North Carolina 27708, USA

Extracellular Ca^{2+} (Ca_o^{2+}) is required for various physiological and developmental processes in animals and plants^{1–3}. In response to varied Ca_o^{2+} levels, plants maintain relatively constant internal Ca^{2+} content, suggesting a precise regulatory mechanism for Ca^{2+} homeostasis⁴. However, little is known about how plants monitor Ca_o^{2+} status and whether Ca_o^{2+} -sensing receptors exist. The effects of Ca_o^{2+} on guard cells in promoting stomatal closure by inducing increases in the concentration of cytosolic Ca^{2+} ($[\text{Ca}^{2+}]_i$)^{5–8} provide a clue to Ca_o^{2+} sensing. Here we have used a functional screening assay in mammalian cells⁹ to isolate an *Arabidopsis* complementary DNA clone encoding a Ca^{2+} -sensing receptor, CAS. CAS is localized to the plasma membrane, exhibits low-affinity/high-capacity Ca^{2+} binding, and mediates Ca_o^{2+} -induced $[\text{Ca}^{2+}]_i$ increases. CAS is expressed predominantly in the shoot, including guard cells. Repression of CAS disrupts Ca_o^{2+} signalling in guard cells, and impairs bolting (swift upward growth at the transition to seed production) in response to Ca^{2+} deficiency, so we conclude that CAS may be a primary transducer of Ca_o^{2+} in plants.

In plants, a high proportion of the total Ca^{2+} is often located in the cell wall and at the exterior surface of the plasma membrane^{2,3,10}. Ca_o^{2+} is important for the stability of the wall and the membrane, and is essential in many physiological processes, such as root-tip growth, pollen-tube elongation, light perception, and phytohormone action². In contrast to cytosolic Ca^{2+} sensing^{3,11}, how plants perceive Ca_o^{2+} and coordinate their growth and development with the status of Ca_o^{2+} remains largely unknown. It has long been established that Ca_o^{2+} promotes stomatal closure^{5,12} by triggering

guard-cell $[Ca^{2+}]_i$ increases and oscillations in *Commelina*⁶. We reasoned that Ca_0^{2+} -induced $[Ca^{2+}]_i$ increases (CICI) in *Commelina* guard cells might represent a mechanism by which plants sense Ca_0^{2+} . To investigate the molecular nature of CICI, we re-examined the Ca_0^{2+} effect in *Arabidopsis* guard cells using cameleon-based Ca^{2+} -imaging techniques⁷. The addition of $CaCl_2$ induced $[Ca^{2+}]_i$ increases in guard cells (Supplementary Fig. S-1a), consistent with previous studies^{6,7}. It is possible that increased Ca_0^{2+} provides a driving force for Ca^{2+} influx, triggers a signalling cascade for Ca^{2+} release from internal stores, or does both. In both animals and plants, the inositol-1,4,5-trisphosphate (IP_3) pathway is responsible for stimulus-triggered Ca^{2+} release in many instances^{3,13}. Preliminary experiments show that inhibition of phospholipase C, a component of the IP_3 pathway, abolished guard-cell CICI and stomatal closure (Supplementary Fig. S-1), indicating that CICI may involve Ca_0^{2+} sensing by a receptor (CAS).

To study the putative CAS further, we attempted to clone CAS from *Arabidopsis* using a functional screening approach, as described for the capsaicin receptor⁹. Messenger RNA from leaves was used to construct a cDNA library, which was subdivided into pools. Each pool was transiently transfected into human embryonic kidney (HEK293) cells. Transfected cells were screened for CICI using ratiometric imaging of the fluorescent Ca^{2+} -sensitive dye Fura-2. Ca_0^{2+} at 2.5 mM caused $[Ca^{2+}]_i$ increases in a few cells transfected with empty vector (Fig. 1, top). A positive pool, pool-58, was identified, which resulted in more Ca_0^{2+} -sensitive cells (Fig. 1, middle). This pool was divided into ten sub-pools and re-screened. After subdividing the pool several times, a single 1.4-kb cDNA was isolated that conferred Ca_0^{2+} sensitivity to transfected HEK293 cells (Fig. 1, bottom), and was thus named CAS.

The CAS cDNA encodes a protein of 387 amino acids with a calculated molecular mass of 41,268 Da (Fig. 2a). CAS is a single-copy gene in the *Arabidopsis* genome. Database searches identified several homologues in other plant species but not in animals.

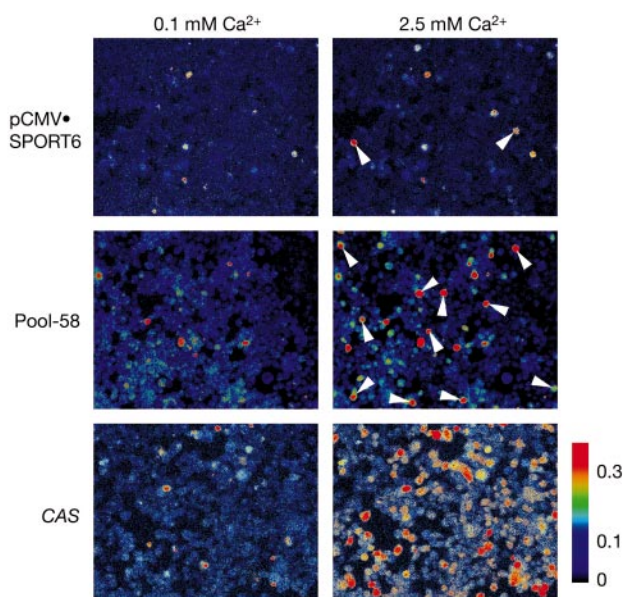


Figure 1 Calcium-imaging-based expression cloning of a Ca_0^{2+} -sensing receptor from an *Arabidopsis* cDNA library. HEK293 cells were transiently transfected with empty pCMV•SPORT6 vector, Pool-58 or CAS, and the $[Ca^{2+}]_i$ status before (0.1 mM) and after treatment with 2.5 mM Ca_0^{2+} was analysed using Fura-2-based Ca^{2+} imaging⁹. $[Ca^{2+}]_i$ increases are indicated by increased Fura-2 emission ratios ($F_{340\text{ nm}}/F_{380\text{ nm}}$) and scaled by a pseudo-colour bar in relative units. Arrows show cells with CICI.

Hydrophobicity analysis showed that CAS has a single transmembrane domain (Fig. 2a, b). The carboxy terminus is possibly intracellular, and has sequence similarity to a rhodanese domain, which is likely to have a role in protein–protein interactions. The putative extracellular amino terminus has no significant homology to any known functional domains. It should be noted that in contrast to cytosolic Ca^{2+} sensors, such as calmodulin (CaM)¹¹, CAS is unlikely to contain high-affinity Ca^{2+} -binding sites, such as EF-hands, as physiological Ca_0^{2+} is in the submillimolar range^{2,14}. Nevertheless, acidic amino acids have been reported to act as low-affinity Ca^{2+} -binding sites in animal cells¹⁵. The N terminus contains several acidic amino acids and its isoelectric point is ~ 5.3 (Fig. 2a).

To test whether CAS binds to Ca^{2+} , we prepared recombinant

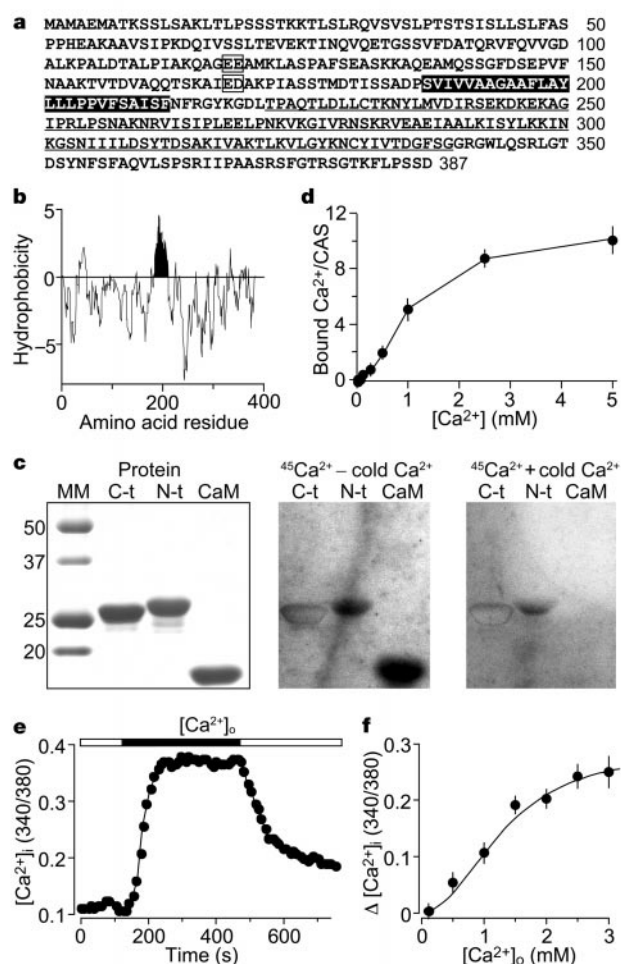


Figure 2 CAS encodes a low-affinity/high-capacity Ca^{2+} -binding protein. **a**, Deduced amino acid sequence encoded by the CAS cDNA. Filled and open boxes show the predicted single transmembrane domain and putative Ca^{2+} -binding sites, respectively. The rhodanese homology domain is underlined. **b**, Hydrophobicity analysis of CAS. **c**, ^{45}Ca blot analysis of the CAS N and C termini (N-t and C-t). Purified recombinant N terminus, C terminus and bovine calmodulin (CaM) were loaded, stained with Coomassie blue (left), and probed with ^{45}Ca in the absence (middle) or the presence of cold $CaCl_2$ (5 mM, right). MM, molecular marker (kDa). **d**, Low-affinity/high-capacity Ca^{2+} binding to the N terminus measured by equilibrium dialysis. **e**, Time-course analysis of CAS activation and inactivation in CAS-transfected HEK293 cells. Open and filled bars show the Ca_0^{2+} concentrations of 0.1 and 2.5 mM, respectively. **f**, Averaged increases in relative $[Ca^{2+}]_i$ ($\Delta F_{340\text{ nm}}/F_{380\text{ nm}}$) plotted as a function of applied $[Ca^{2+}]_o$ from experiments as in **e** ($n = 20$ cells). Data were fitted to a Hill equation.

proteins for both the N terminus and the C terminus (Fig. 2c, left), and assayed their abilities to bind Ca^{2+} . When blotting with ^{45}Ca only, the N terminus and CaM bound to ^{45}Ca , but not the C terminus (Fig. 2c, middle). When blotting with ^{45}Ca and cold CaCl_2 , only the N terminus bound to ^{45}Ca (Fig. 2c, right), showing qualitatively that the N terminus is indeed a Ca^{2+} -binding domain, and contains low-affinity/high-capacity Ca^{2+} -binding sites compared with CaM. To analyse the Ca^{2+} -binding quantitatively, we conducted equilibrium dialysis using the purified N-terminal protein. Figure 2d shows that the N terminus bound 10 to 12 Ca^{2+} per CAS molecule with a dissociation constant $K_d \approx 1.2$ mM, confirming the ^{45}Ca -binding data. Furthermore, we analysed the kinetics of activation and inactivation of CAS in HEK293 cells, which correlate with the association and dissociation between CAS and Ca_o^{2+} , respectively. The activation time course could be described by a simple sigmoid equation¹⁶, with an activation time-constant of 28.6 s (Fig. 2e). A triple exponential decay function could be fitted¹⁶ to the inactivation of CAS (Fig. 3e) with an inactivation time-constant of 18.2 s. A Hill curve could be fitted¹⁶ to the data of CAS activation (Fig. 2f) with a K_d of ~ 1.5 mM and a Hill coefficient of ~ 2 , consistent with the K_d and multiple Ca_o^{2+} -binding sites observed using equilibrium dialysis. Taken together, these data demonstrate that CAS contains low-affinity/high-capacity Ca^{2+} -binding sites on the N terminus.

Subsequently, we determined the expression and subcellular localization of CAS in *Arabidopsis*. RNA blot analysis shows that CAS is predominantly expressed in shoots (Fig. 3a). A *CAS promoter::GUS* fusion shows that CAS is expressed in guard cells (Fig. 3b). We expressed 35S::CAS-GFP and 35S::GFP in onion epidermal cells, and found that CAS-GFP (green fluorescent protein) was exclusively localized to the cell surface, whereas GFP alone was localized throughout the cells (Fig. 3c). After the epidermal cells were incubated in 0.8 M mannitol, the fusion proteins and the plasma membrane retracted from the cell wall due to plasmolysis (not shown). We also found CAS in the membrane fraction but not in the soluble fraction of proteins from leaves (Fig. 3d). Finally, we confirmed the guard-cell surface localization of CAS using immunolocalization¹⁷ (Fig. 3e).

To assess the function of CAS in plants, first we examined whether CAS is required for guard-cell Ca_o^{2+} signalling. As we were unable to find CAS-null lines, we introduced a 35S::CAS antisense construct into *Arabidopsis*. CAS expression was dramatically reduced at the mRNA and protein levels in all six antisense lines tested (AS1 to AS6) with three of them shown (Fig. 4a, b). To analyse CICI in guard cells, we introduced aameleon construct into AS1 and AS2 plants, and found that the CICI was disrupted significantly in these lines compared with the wild type (WT; Fig. 4c, d). Ca_o^{2+} -induced stomatal closure was also abolished in these lines (Fig. 4e). However, ABA-induced stomatal closure was virtually unaffected (not shown), suggesting that CAS antisense specifically knocked out Ca_o^{2+} signalling, and that CAS has an essential role in guard-cell Ca_o^{2+} signalling. Note that, apart from Ca_o^{2+} , stomatal movements are regulated by various abiotic and biotic factors^{5,8,18}. Apparently, guard cells must integrate several factors to control stomatal apertures appropriately^{8,18}; CAS, several Ca^{2+} influx and release mechanisms, and Ca^{2+} pumps must act in concert to regulate cytosolic free Ca^{2+} status^{3,8}.

Calcium is an abundant and essential divalent cation nutrient in higher plants, and has a central role in numerous physiological and developmental processes^{2,3,10,14}. To examine whether CAS also functions in processes other than stomatal closure, we analysed plant growth and development in response to limited Ca^{2+} supplies. Figure 4f shows that AS1 plants showed a severe defect in bolting in response to Ca^{2+} deficit compared with the wild type. AS1 plants bolted eventually at reduced Ca^{2+} but much later than the wild type (Fig. 4g). In contrast to wild-type plants, AS1 plants could not flower and displayed severe symptoms of Ca^{2+} deficiency (not

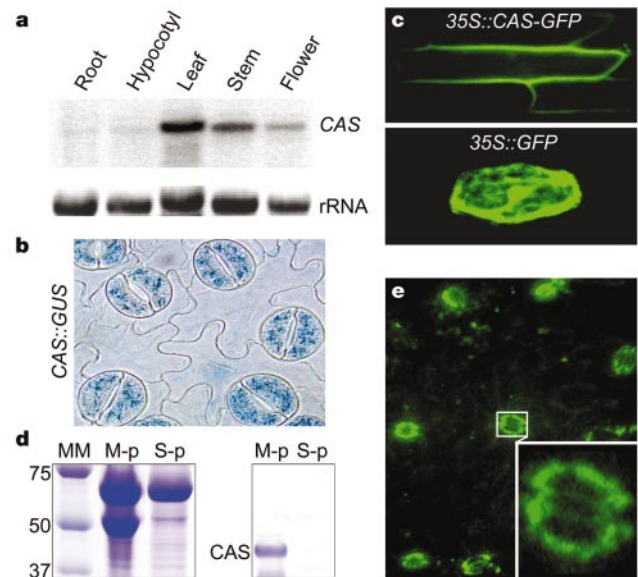


Figure 3 Expression pattern and subcellular location of CAS. **a**, Northern blot analysis of CAS expression in *Arabidopsis*. Ribosomal RNA was used as a loading control. **b**, *CAS promoter::GUS* expression in guard cells. **c**, Plasma membrane localization of CAS-GFP fusion protein in onion epidermal cells transiently expressing a 35S::CAS-GFP construct with the 35S::GFP as a control. **d**, Membrane-associated (M-p) and soluble protein (S-p) from leaves was loaded, stained with Coomassie blue (left), and blotted with anti-CAS antibodies (right). **e**, CAS protein localizes to the cell surface of guard cells. CAS protein was detected using anti-CAS antibodies. Inset shows one enlarged stoma.

shown). The Ca^{2+} content of these plants was further analysed and no significant differences between AS1 and wild-type plants were found under the imposed growth conditions ($\sim 100\%$ humidity, not shown). Similar phenotypes were observed in AS2 to AS6 plants. Bolting represents an important developmental switch in flowering plants¹⁹. It is possible that a threshold of Ca_o^{2+} has to be reached and sensed by CAS to signal the switch to bolting. In the antisense plants, although the Ca_o^{2+} level is not changed, the malfunction of Ca_o^{2+} sensing cannot sensitively detect the level of Ca_o^{2+} and signals insufficient Ca_o^{2+} for bolting, suggesting that CAS has a key role in *Arabidopsis* development.

We have cloned and characterized a cell surface Ca_o^{2+} -sensing receptor in *Arabidopsis*. We have also identified the essential role of CAS in Ca_o^{2+} signalling in guard cells and bolting. Molecular genetic, cell biological, and physiological analyses lead us to propose that Ca_o^{2+} activates CAS, which converts the Ca_o^{2+} signal into cytosolic free Ca^{2+} . It is possible that CAS may initiate the IP_3 pathway in *Arabidopsis*, as suggested by our preliminary observation (Supplementary Fig. S-1), consistent with IP_3 -induced stomatal closure²⁰. In contrast, previous studies have shown that the CICI in *Commelina* guard cells is not affected by inhibition of phospholipase C, but by inhibition of Ca^{2+} channels^{6,21}, suggesting that Ca_o^{2+} triggers Ca^{2+} influx. Therefore, it remains to be determined which Ca^{2+} -mobilizing machineries—such as Ca^{2+} releases triggered by IP_3 , cyclic ADP-ribose or nicotinic acid adenine dinucleotide phosphate, or Ca^{2+} influx channels—are activated by CAS. Functional analysis of CAS-mediated CICI in heterologous expression systems and in plants, and identification of the downstream components of CAS, should help to clarify these issues.

The CAS-mediated Ca_o^{2+} -induced stomatal closure is possibly physiologically relevant. The distribution of Ca^{2+} in plants is affected by the rate of transpiration, which is predominantly controlled by stomata^{5,10}. Calcium moves in relatively large amounts to highly transpiring tissues, but much less moves to weakly

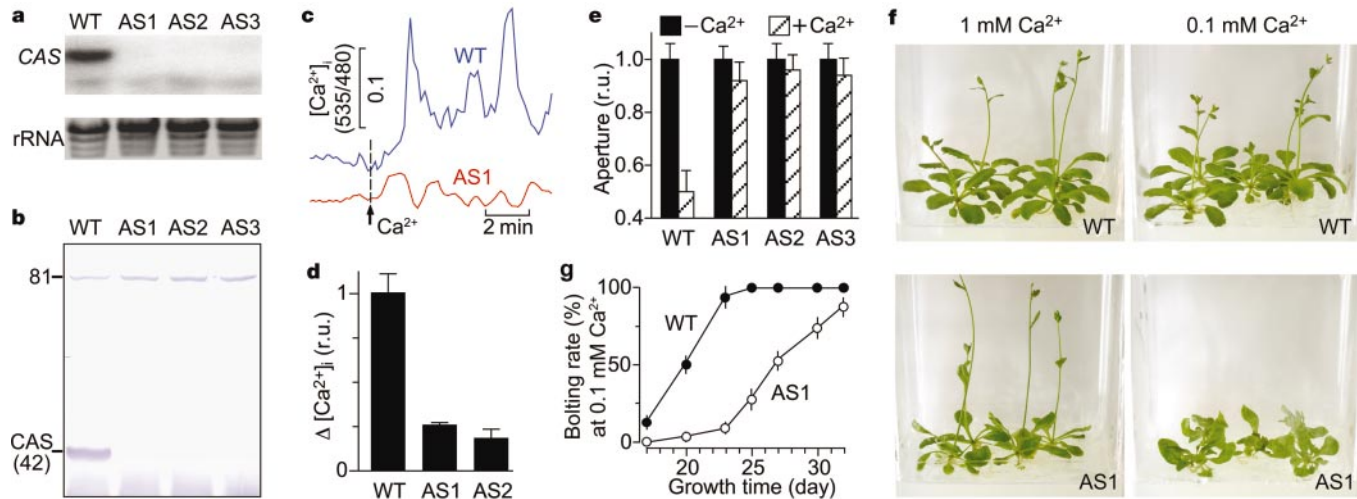


Figure 4 CAS is required for guard-cell Ca^{2+} signalling and plant development. **a**, Northern blot analysis of CAS mRNA levels in leaves from wild-type and CAS antisense transgenic plants (AS1 to AS3). **b**, Western blot analysis of CAS protein levels using total proteins from leaves. An 81-kDa band stained non-specifically by anti-CAS antibodies. **c**, Defect in guard-cell Cl^- in AS1 plants. AS1 was transformed with a cameleon construct carrying a basta resistance marker. 2 mM Ca^{2+} was added as described. Other conditions were the same as in Supplementary Fig. S-1. **d**, Increase in guard-cell $[\text{Ca}^{2+}]_i$ ($F_{535\text{ nm}}/F_{480\text{ nm}}$) in response to Ca^{2+} treatment in the wild type, AS1 and AS2 from

experiments as in **c** ($n = 48$ to 56 guard cells). $[\text{Ca}^{2+}]_i$ increases were normalized to that of the wild type. r.u., relative unit. **e**, Defects in Ca^{2+} -induced stomatal closure in CAS antisense lines. Stomatal apertures were normalized to those without 2 mM Ca^{2+} treatment ($n = 120$ stomata). **f**, CAS is required for bolting in response to reduced Ca^{2+} supplies. Both wild-type and AS1 plants were grown in media containing 1 mM (left) or 0.1 mM Ca^{2+} (right), and plant photos were taken after 23 days of growth. **g**, Quantitative analysis of the late-bolting phenotype of AS1 grown at 0.1 mM Ca^{2+} . Data from four separate experiments and total of ~60 seedlings are shown.

transpiring tissues¹⁰. From the standpoint of Ca^{2+} supply, plants must downregulate transpiration when the Ca^{2+} -unloading rate is high, and vice versa. It is likely that excessive water evaporation from stomatal pores results in an increase in the local concentration of Ca^{2+} outside guard cells to submillimolar or millimolar. Consequently, the elevated Ca^{2+} activates CAS, which mediates stomatal closure, preventing excessive Ca^{2+} unloading. Thus, Ca^{2+} -induced stomatal closure could function as a feedback mechanism of Ca^{2+} supply. Interestingly enough, our preliminary data show that the Ca^{2+} content was ~40% higher in antisense plants than in the wild type when they were grown in soil, supporting the feedback hypothesis. Precise information on the concentration range of Ca^{2+} is lacking¹⁴, and data taken from the literature vary from 10 to 1,000 μM (refs 14, 22). It is especially difficult to precisely detect Ca^{2+} outside guard cells, as water evaporation from stomatal pores may cause uneven distribution of apoplastic solutes, including Ca^{2+} . Our finding that CAS is required for guard-cell Ca^{2+} signalling and bolting suggests that CAS has a central role in diverse Ca^{2+} -dependent processes. The cloning of CAS and manipulation of its activity now make it possible to assess the molecular function of Ca^{2+} in plants. □

Methods

Imaging of $[\text{Ca}^{2+}]_i$ in guard cells

Leaf epidermal peels from soil-grown *Arabidopsis thaliana* (Columbia) were placed in a microwell chamber containing 50 μM CaCl_2 , 5 mM KCl, 10 mM MES-Tris, pH 6.15. Cameleon-based Ca^{2+} imaging was performed as described⁷ using a fluorescence microscope (Axiovert 200; Zeiss) equipped with two filter wheels (Lambda 10-2; Sutter), and a cooled CCD camera (CoolSNAP_{FX}; Roper). Excitation was provided at 440 nm, and emission ratiometric ($F_{535\text{ nm}}/F_{480\text{ nm}}$) images were collected using Imaging Workbench 4.0 software (Axon).

Stomatal aperture bioassay

Rosette leaves were floated in solutions containing 50 μM CaCl_2 , 10 mM KCl, 10 mM MES-Tris, pH 6.15, for 2 h. CaCl_2 at 2 mM was added to the solutions for 2 h to assay stomatal closure as described¹⁸.

Expression cloning

mRNAs were isolated from *Arabidopsis* leaves. cDNAs were synthesized, and ligated into the *SalI* and *NorI* sites of pCMV-SPORT6 vector (Invitrogen). Recombinant plasmids were introduced into *Escherichia coli*. The resulting 2×10^6 clones were divided into 100 pools for transfection of HEK293 cells. HEK293 cells were grown and maintained in DMEM medium supplemented with 10% fetal bovine serum, 0.1% penicillin and streptomycin. For transfection, cells were seeded onto eight-well-chambered coverglasses (Nunc) for 24 h, and transfected with plasmid DNA from individual pools using LipofectAMINE 2000 reagent (Invitrogen). Cells were loaded with Fura-2-AM (5 μM ; Molecular Probes), and a Ca^{2+} -imaging assay was performed in the HEK293 cells as described⁹.

Ca^{2+} -binding assays

For ⁴⁵Ca blot analysis²³, the N terminus (1–180) and C terminus (211–386) were subcloned into pET-30a and expressed in *E. coli*. The purified N and C termini together with bovine CaM at 10 μg were separated by SDS-PAGE, and transferred to membranes (Millipore). The membranes were washed in Ca^{2+} -free solution containing 5 mM MgCl_2 , 60 mM KCl, 10 mM MES-KOH, pH 6.5, blotted with 2 $\mu\text{Ci ml}^{-1}$ (3 μM) fresh ⁴⁵Ca (<14 days; PerkinElmer, 370 MBq ml^{-1}) for 10 min, rinsed for 5 min in 50% ethanol, and exposed to X-ray films. Unlabelled Ca^{2+} at 5 mM was added as indicated. Equilibrium dialysis was performed using dialysis cells (100 μl) separated by dialysis membranes (Bel-Art) as described²⁴. Purified N terminus (66 μM) was dialysed in a solution containing 0.1 μCi ⁴⁵Ca (1.5 μM), 0–10 mM CaCl_2 , 100 mM KCl, 10 mM MOPS, pH 7.5, for 24 h. Radioactivity was determined by liquid scintillation counting.

Expression and localization analysis

Northern blot analysis, histochemical staining for GUS activity and immunolocalization were performed as described^{17,25}. For northern blot, ~15 μg RNA was loaded and blotted with a ³²P-labelled CAS probe. For GUS staining, a genomic fragment of 1.5 kb in the promoter region of CAS was amplified by PCR with primers: 5'-GCCCGCGGTGTTT GTTGTATTGTTTGGGTA and 5'-GGCTCGAGGGTTCTCTCTGCCACACTT, and cloned into a GUS binary vector pBI101. Transgenic plants were generated using *Agrobacterium*-mediated transformation and selected by kanamycin²⁶. Independent homozygous transformants carrying a single insertion in the T3 generation were analysed further. CAS antisense plants were generated similarly. For GFP-based analysis, 35S::CAS-GFP or 35S::GFP was transiently expressed in onion epidermal cells by bombardment²⁵. For immunolocalization, polyclonal antibodies were raised in rabbits using the purified N and C termini as the antigen. The leaf epidermal peels were placed on poly-L-lysine coated slides. Affinity-purified primary anti-CAS and fluorescein isothiocyanate-conjugated anti-rabbit secondary antibodies (Sigma) were diluted 1:10 and 1:200, respectively.

Western blot

Total proteins (10 μg) from *Arabidopsis* leaves were separated on SDS-PAGE, and incubated with anti-CAS antibody followed by alkaline phosphatase-conjugated

anti-rabbit antibody (Sigma), and developed using NBT/BCIP substrate. The leaf homogenate was centrifuged at 13,000g, and the supernatant was further centrifuged at 80,000g to separate the soluble proteins (supernatant) and membrane-associated proteins (pellet).

Ca²⁺-deficiency and Ca²⁺-content analysis

Wild-type and antisense plants were grown in the MS-based media containing MS minor salts (Sigma), 1% sucrose (Sigma), 1% agar (Becton Dickinson) and MS major salts (Sigma) with Ca²⁺ supplemented to 0.1 mM or 1 mM CaCl₂. Each pot contained 50 ml media. For Ca²⁺-content analysis, shoots were digested with 70% nitric acid, and the total Ca²⁺ content was determined using an atomic absorption spectrometer. Data were analysed and presented as described^{16,18}.

Received 8 February; accepted 18 July 2003; doi:10.1038/nature01932.

1. Brown, E. M. & MacLeod, R. J. Extracellular calcium sensing and extracellular calcium signaling. *Physiol. Rev.* **81**, 239–297 (2001).
2. Hepler, P. K. & Wayne, R. O. Calcium and plant development. *Annu. Rev. Plant Physiol.* **36**, 397–439 (1985).
3. Sanders, D., Pelloux, J., Brownlee, C. & Harper, J. F. Calcium at the crossroads of signaling. *Plant Cell* **14**, S401–S417 (2002).
4. Loneragan, J. F. & Snowball, K. Calcium requirements of plants. *Aust. J. Agr. Res.* **20**, 465–478 (1969).
5. MacRobbie, E. Calcium and ABA-induced stomatal closure. *Phil. Trans. R. Soc. Lond. B* **338**, 5–18 (1992).
6. McAinsh, M. R., Webb, A. A. R., Taylor, J. E. & Hetherington, A. M. Stimulus-induced oscillations in guard cell cytosolic free calcium. *Plant Cell* **7**, 1207–1219 (1995).
7. Allen, G. J. *et al.* A defined range of guard cell calcium oscillation parameters encodes stomatal movements. *Nature* **411**, 1053–1057 (2001).
8. Schroeder, J. I., Kwak, J. M. & Allen, G. J. Guard cell abscisic acid signalling and engineering drought hardiness in plants. *Nature* **410**, 327–330 (2001).
9. Caterina, M. J. *et al.* The capsaicin receptor: a heat-activated ion channel in the pain pathway. *Nature* **389**, 816–824 (1997).
10. Clarkson, D. T. Calcium transport between tissues and its distribution in the plant. *Plant Cell Environ.* **7**, 449–456 (1984).
11. Luan, S., Kudla, J., Rodriguez-Concepcion, M., Yalovsky, S. & Gruissem, W. Calmodulins and calcineurin B-like proteins: calcium sensors for specific signal response coupling in plants. *Plant Cell* **14**, S389–S400 (2002).
12. Schwartz, A. Role of Ca²⁺ and EGTA on stomatal movements in *Commelina communis* L. *Plant Physiol.* **79**, 1003–1005 (1985).
13. Berridge, M. J., Bootman, M. D. & Lipp, P. Calcium—a life and death signal. *Nature* **395**, 645–648 (1998).
14. Sattelmacher, B. The apoplast and its significance for plant mineral nutrition. *New Phytol.* **149**, 167–192 (2001).
15. Chapman, E. R. Synaptotagmin: a Ca²⁺ sensor that triggers exocytosis? *Nature Rev. Mol. Cell Biol.* **3**, 498–508 (2002).
16. Pei, Z.-M., Baizabal-Aguirre, V. M., Allen, G. J. & Schroeder, J. I. A transient outward-rectifying K⁺ channel current down-regulated by cytosolic Ca²⁺ in *Arabidopsis thaliana* guard cells. *Proc. Natl Acad. Sci. USA* **95**, 6548–6553 (1998).
17. Friml, J., Wisniewska, J., Benkova, E., Mendgen, K. & Palme, K. Lateral relocation of auxin efflux regulator PIN3 mediates tropism in *Arabidopsis*. *Nature* **415**, 806–809 (2002).
18. Pei, Z.-M. *et al.* Calcium channels activated by hydrogen peroxide mediate abscisic acid signalling in guard cells. *Nature* **406**, 731–734 (2000).
19. Simpson, G. G. & Dean, C. *Arabidopsis*, the Rosetta stone of flowering time? *Science* **296**, 285–289 (2002).
20. Gilroy, S., Read, N. D. & Trewavas, A. J. Elevation of cytoplasmic Ca²⁺ by caged calcium or caged inositol triphosphate initiates stomatal closure. *Nature* **346**, 769–771 (1990).
21. Staxen, I. *et al.* Abscisic acid induces oscillations in guard-cell cytosolic free calcium that involve phosphoinositide-specific phospholipase C. *Proc. Natl Acad. Sci. USA* **96**, 1779–1784 (1999).
22. Roelfsema, M. R. G. & Hedrich, R. Studying guard cells in the intact plant: modulation of stomatal movement by apoplastic factors. *New Phytol.* **153**, 425–431 (2002).
23. Maruyama, K., Mikawa, T. & Ebashi, S. Detection of calcium binding proteins by ⁴⁵Ca autoradiography on nitrocellulose membrane after sodium dodecyl sulfate gel electrophoresis. *J. Biochem.* **95**, 511–519 (1984).
24. Wakabayashi, H., Schmidt, K. M. & Fay, P. J. Ca²⁺ binding to both the heavy and light chains of factor VIII is required for cofactor activity. *Biochemistry* **41**, 8485–8492 (2002).
25. Li, X., Zhang, Y., Clarke, J. D., Li, Y. & Dong, X. Identification and cloning of a negative regulator of systemic acquired resistance, SN1, through a screen for suppressors of *npr1-1*. *Cell* **98**, 329–339 (1999).
26. Clough, S. J. & Bent, A. F. Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J.* **16**, 735–743 (1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank L. Jing, Y. Hao and Y. He for assistance; W. Zhang for HEK293 cells and discussions on expression cloning; D. McClay and R. Fehon for use of cell culture and imaging facilities; J. Schroeder, J. Kwak, J. Harper and A. VanDogen for vectors; J. Schroeder, T. Sun and R. Venters for discussions; and J. Siedow, X. Dong, W. Durrant, P. Benfey, W. Zhang and R. Fehon for comments on the manuscript. R.T. was supported in part by a Hargitt Fellowship. This work was supported by start-up funds from Duke University and by a grant from the NSF to Z.-M.P.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to Z.-M.P. (zpei@duke.edu). The GenBank accession code for CAS is AY341888.

Structural basis for modulation and agonist specificity of HCN pacemaker channels

William N. Zagotta¹, Nelson B. Olivier², Kevin D. Black¹, Edgar C. Young³, Rich Olson² & Eric Gouaux^{2,4}

¹Department of Physiology and Biophysics, Howard Hughes Medical Institute, Box 357290, University of Washington School of Medicine, Seattle, Washington 98195-7290, USA

²Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street, New York, New York 10032, USA

³Center for Neurobiology and Behavior, Howard Hughes Medical Institute, Columbia University, 722 West 168th Street, New York, New York 10032, USA

⁴Howard Hughes Medical Institute, Columbia University, 650 West 168th Street, New York, New York 10032, USA

The family of hyperpolarization-activated, cyclic nucleotide-modulated (HCN) channels are crucial for a range of electrical signalling, including cardiac and neuronal pacemaker activity, setting resting membrane electrical properties and dendritic integration¹. These nonselective cation channels, underlying the I_f, I_h and I_q currents of heart and nerve cells, are activated by membrane hyperpolarization and modulated by the binding of cyclic nucleotides such as cAMP and cGMP². The cAMP-mediated enhancement of channel activity is largely responsible for the increase in heart rate caused by β-adrenergic agonists³. Here we have investigated the mechanism underlying this modulation by studying a carboxy-terminal fragment of HCN2 containing the cyclic nucleotide-binding domain (CNBD) and the C-linker region that connects the CNBD to the pore. X-ray crystallographic structures of this C-terminal fragment bound to cAMP or cGMP, together with equilibrium sedimentation analysis, identify a tetramerization domain and the mechanism for cyclic nucleotide specificity, and suggest a model for ligand-dependent channel modulation. On the basis of amino acid sequence similarity to HCN channels, the cyclic nucleotide-gated, and eag- and KAT1-related families of channels are probably related to HCN channels in structure and mechanism.

The HCN channels belong to the superfamily of voltage-activated channels (refs 1, 2 and Fig. 1a). They are composed of four subunits arranged around a centrally located pore. The intracellular C-terminal region of each subunit contains a CNBD and C-linker region that connects the CNBD to the pore. The binding of cAMP to the CNBD accelerates the activation kinetics and shifts the voltage dependence of activation to more positive voltages. This modulation causes the channels to open faster and more completely in the presence of cAMP. In vertebrates, the HCN family comprises four members (HCN1–HCN4), which are related by roughly 90% sequence identity and show varying voltage dependence, cAMP-dependent modulation and kinetics of activation. The C-linker and CNBD of HCN2 share amino acid sequence similarity with domains in the cyclic nucleotide-gated (CNG), eag- and KAT1-related families of channels (Fig. 1b). CNG and related channels also show a cyclic nucleotide-dependent increase in open probability^{4–6}.

To understand the structural basis for the modulation of HCN, CNG and related ion channels by cyclic nucleotides, we determined the atomic resolution structures of two closely related C-terminal fragments of HCN2 by X-ray crystallography. The two constructs begin at Asp 443, just after the end of the S6 transmembrane segment, extend through the C-linker and CNBD, and end at either Asn 640 for HCN2I or His 645 for HCN2J (Fig. 1b). Both constructs were crystallized in the presence of cAMP or cGMP and in different

lattices, thus allowing for several views of the proteins in different environments.

The structure of the HCN2I construct in complex with cAMP was solved to 2.0 Å resolution by multiwavelength anomalous diffraction (MAD) using a selenomethionine derivative, and subsequent structures were determined by molecular replacement (Table 1). Electron density was visible for nearly all residues except for amino acids 567–570 in HCN2I and the last few amino acids of each construct. Because the structures of the different constructs in different crystal forms are essentially identical and the longer HCN2J construct contains 12 additional ordered residues, we have focused on the HCN2J structure.

The HCN2 C-terminal region is composed of two domains (Fig. 2a). The C-linker domain consists of six α -helices, designated A' to F', which are separated by short loops. The CNBD follows the C-linker domain and includes four α -helices (A, P, B, C) with a

β -roll between the A- and B-helices. The β -roll comprises eight β -strands in a jelly-roll-like topology. Cyclic nucleotides are bound inside the jelly-roll and interact with the β -roll and the C-helix.

The HCN2 C-terminal regions assemble as tetramers with a four-fold axis of rotational symmetry (Fig. 2b), mirroring the predicted four-fold symmetry of the transmembrane ion channel⁷. The amino termini are close to each other on the top of the structure, poised to connect to the S6 segments, whereas the C termini are on the bottom face. The subunit–subunit interface is large, with each subunit burying about 2,300 Å² of solvent-accessible surface area on tetramer formation, and is formed almost completely by the C-linker domains.

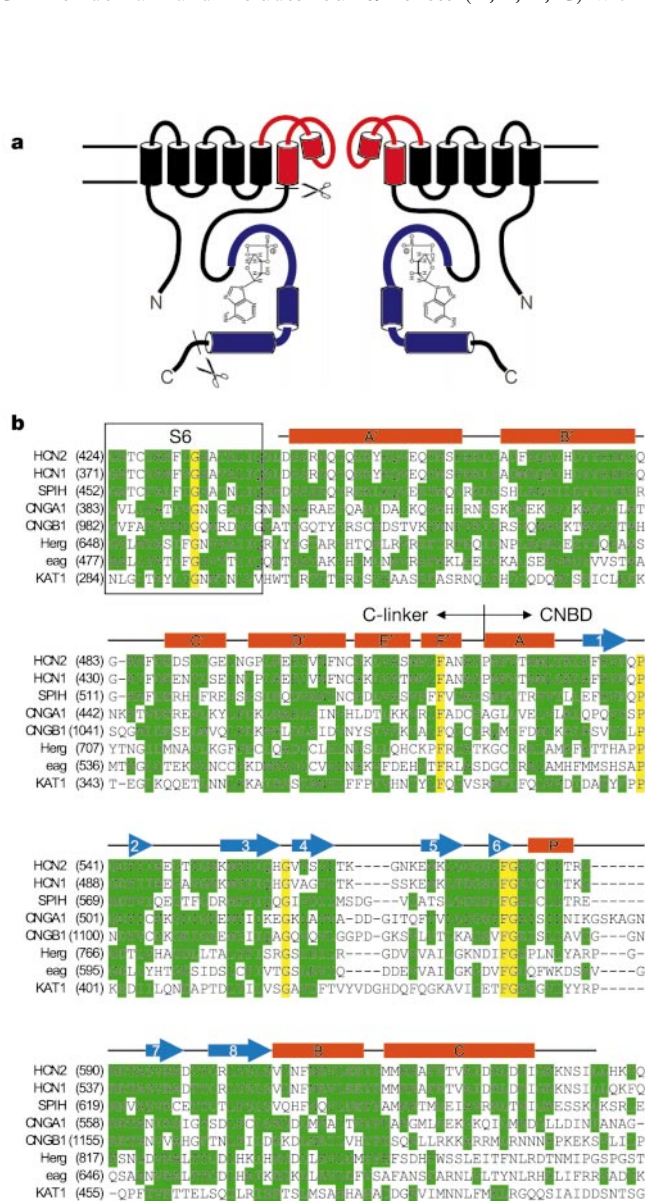


Figure 1 HCN2 channel topology and alignment with related channels. **a**, Two of the four subunits that form the HCN2 channel. The pore and S6 transmembrane segments are red, the cyclic nucleotide-binding domain is blue, and the B- and C-helices are represented as cylinders. Scissors indicate the boundaries of constructs used in this study. **b**, Sequence alignment of the C-terminal region of ion channel proteins related to the mouse HCN2 channel. Red bars and blue arrows denote α -helices and β -strands, respectively. Helices are labelled alphabetically, strands are labelled numerically.

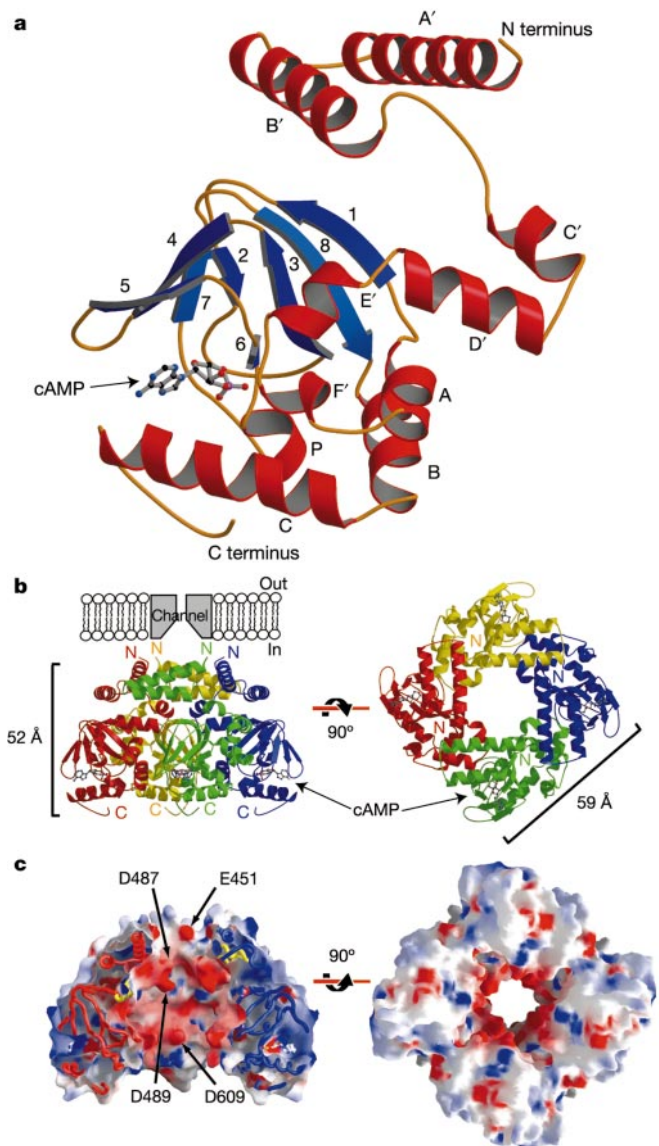


Figure 2 Structure of the mouse HCN2 C-linker and CNBD construct bound to cAMP. **a**, Ribbon representation of a single protomer of HCN2J with cAMP. **b**, HCN2 tetramer viewed perpendicular (left) and parallel (right) to the four-fold axis. Each subunit is shown in a different colour. **c**, Molecular surface representation of the tetramer, coloured according to the calculated electrostatic potential and viewed parallel to the four-fold axis from the intracellular side of the membrane (right). Residues Glu 451, Asp 487, Asp 489 and Asp 609 protrude toward the central axis of the tetramer. The view on the left is perpendicular to the four-fold axis, and the molecule has been sliced along the symmetry axis.

The pyramid-shaped tetramer measures 59 Å at its base and 52 Å in height, and has a prominent hole down the middle that is concentric with the putative transmembrane pore (Fig. 2c). At its narrowest point, the hole is 12 Å in diameter and lined by acidic residues, in particular Asp 487 and Asp 489. Above and below this constriction are two additional rings of negatively charged residues, Glu 451 and Asp 609. The size and polarity of the hole are sufficient to accommodate hydrated cations, although there is no evidence that this hole participates in the ion permeation pathway. In addition, deletion of the whole C terminus of HCN1 does not disrupt ion conduction or voltage gating⁸.

The CNBD of HCN2 has the same fold as the CNBD of other cAMP-binding proteins such as *Escherichia coli* catabolite gene activator protein (CAP)⁹ and the regulatory subunit of cAMP-dependent protein kinase A (PKA; ref. 10, Fig. 2a and Supplementary Information). Like CAP, the CNBD contains an eight-stranded antiparallel β-roll, which is preceded by the A-helix and followed by the B- and C-helices. An additional α-helix, the P-helix, is located between β-strands 6 and 7, as in the structure of PKA¹⁰.

Whereas in CAP the subunits dimerize through a coiled-coil interaction of their C-helices, in HCN2 the CNBDs do not interact appreciably. cAMP binds in the anti configuration between the β-roll and the C-helix (Fig. 3a, b). The ribose and cyclic phosphate interact exclusively with the β-roll, primarily through an electrostatic interaction between Arg 591 and the phosphate of cAMP and through hydrogen-bonding interactions with the oxygens on the phosphoribose.

Consistent with this structure, mutations of Arg 591 have been shown to cause a large decrease in HCN2 affinity for cAMP¹¹. The purine ring interacts with both the β-roll and the C-helix (Fig. 3c). The interactions with the C-helix are of particular note because mutations in the C-helix in HCN2 and CNG channels selectively perturb the ability of bound ligand to promote opening of the channel^{8,12,13}. For example, Ile 636 forms a hydrophobic interaction with the purine ring of cAMP. In CAP, Thr 128 occupies the same relative position as Ile 636 in HCN2. CAP Thr 128 forms a hydrogen bond with the N6 amine of cAMP and is thought to underlie the cAMP specificity of CAP⁹. In HCN2, the N6 amine instead forms a hydrogen bond with the backbone carbonyl oxygen of Arg 632. Additional interactions with the purine ring are made by the methylene groups of Arg 632 in the C-helix and Val 564, Met 572 and Leu 574 in the β-roll (Fig. 3c).

Although cAMP is the most effective ligand, the HCN channel in the sinoatrial node of the heart is also modulated by cGMP, albeit with a roughly 30-fold lower sensitivity^{3,14}. To determine the cyclic nucleotide specificity of our fragment of HCN2 in the context of the whole channel, we expressed a truncated HCN2 channel with a C-terminal deletion identical to our HCN2J fragment in *Xenopus* oocytes. Inside-out patch-clamp recordings from these channels showed channel behaviour very similar to that of wild-type HCN2 channels (refs 8, 15, and Fig. 3d). The channels were activated by membrane hyperpolarization with a voltage for half-maximal activation ($V_{1/2}$) of -136.2 ± 5.1 mV. cAMP caused a depolarizing shift of 19.2 ± 0.2 mV in the activation voltage with an agonist concentration eliciting a half-maximal shift ($K_{1/2}$) of

0.84 ± 0.03 μM. As with HCN channels in the sinoatrial node, cGMP caused a similar shift in the activation voltage (18.6 ± 1.8 mV) but required tenfold higher concentrations to produce an effect ($K_{1/2} = 8.3 \pm 2.1$ μM).

To determine the mechanism for this nucleotide specificity, we used molecular replacement to solve the crystal structure of the HCN2J fragment in the presence of cGMP at 1.9 Å resolution. Unexpectedly, cGMP binds in the syn conformation, in contrast to the anti configuration seen for cAMP (Fig. 3b). The syn conformation of cGMP allows its purine ring N2 atom to hydrogen bond with Thr 592 of the β-roll (Fig. 3c). Such an interaction has been proposed to account partially for the cGMP specificity of cGMP-dependent protein kinase and CNG channels^{16–18}. The residue at the homologous position in SPIH, a sea urchin HCN channel, is a valine (Fig. 1b), which explains why SPIH is not activated by cGMP¹⁹. Notably, even though this cGMP-specific interaction occurs in HCN2 channels, the channels are still cAMP-specific (Fig. 3d). This cyclic nucleotide specificity probably results from hydrogen bonding between Arg 632 and the N6 amine of cAMP, as well as from hydrophobic interactions between the adenine ring and the binding pocket.

The C-linker domain, which is unrelated to structures currently in the protein data bank, is a crucial portion of HCN2. Not only does it couple the ion channel gating machinery to the ligand-binding domain, but it also mediates most of the subunit–subunit contacts in the tetramer. The first two helices of each subunit (A' and B') form an antiparallel helix–turn–helix motif that interacts with the C' and D' helices of the neighbouring subunit (Fig. 4a), like an elbow resting on the shoulder of its neighbour. This subunit interface is partially hydrophobic, containing seven aromatic and six leucine or isoleucine residues. In addition, there are several hydrogen-bonding interactions between three tyrosines in the B'-helix and residues in the C'-helix of the neighbouring subunit, and a salt bridge between Lys 472 of the B'-helix and Glu 502 of the D'-helix of the neighbouring subunit.

The C-terminal HCN2 fragment also forms a tetramer in solution in the presence of cAMP. Equilibrium sedimentation showed that the ligand-bound fragment exists in a monomer–tetramer equilibrium with an association constant of $3.0 \times 10^{12} \text{ M}^{-3}$ (Fig. 4b). Because the cAMP-bound C-terminal fragment forms a tetramer in solution when it is not tethered to a tetrameric ion channel, it is likely that in the intact ion channel the C-linker and CNBD also form a tetrameric complex, similar to what we have observed in our crystal structures. In the absence of cAMP, the tetrameric association is significantly weaker (Fig. 4b) and the C-terminal fragment is primarily monomeric, with a monomer–tetramer association constant of $1.7 \times 10^{11} \text{ M}^{-3}$ and a monomer–dimer association constant of $4.4 \times 10^3 \text{ M}^{-1}$. The presence of dimer in solution may reflect a 'dimer of dimers' symmetry in the absence of ligand, as has been suggested by functional experiments^{20,21}. Alternatively, it may simply reflect the nonphysiological dimer contacts seen in the crystals.

For the HCN2 channel, cAMP binding to the CNBD destabilizes the closed state relative to the open state of the channel by about 3 kcal mol⁻¹. Previous studies have suggested that the C-terminal

Table 1 Crystallographic refinement statistics

Complex*	Space group	Resolution (Å)	R_{work} † (%)	R_{free} † (%)	Overall (B value)	r.m.s. deviation	
						Bonds (Å)	Angles (°)
J + cAMP	P4 ₂ 2	20–2.3	21.6	26.1	23.7	0.006	1.19
J + cGMP	I4	30–1.9	20.9	23.9	27.3	0.007	1.23
I + cAMP	I4	20–2.0	19.5	24.2	19.8	0.006	1.16

*I, HCN2I construct from Asp 443 to Asn 640; J, HCN2J construct from Asp 443 to His 645.

† $R_{\text{work}} = \sum |F_o| - |F_c| / \sum |F_o|$, where F_o and F_c are observed and calculated structure factors, respectively; 10% of the data was set aside for calculating R_{free} .

region of HCN2 is a negative allosteric modulator of channel opening and that cAMP binding removes this inhibition^{8,15}. In particular, deletion of the CNBD of HCN2 accelerates the activation kinetics and shifts the voltage dependence of activation to more positive voltages, similar to the binding of cAMP⁸. Chimaeras between the HCN1 and HCN2 channels localized much of the difference in cAMP-dependent modulation to differences in the C-linker region¹⁵. A C-linker region in all four subunits is required for cyclic nucleotide-dependent modulation²¹. Together with those results, our data suggest that cAMP binding to the CNBD induces a significant rearrangement of the tertiary structure of the elbow–shoulder interactions and hence the quaternary structure of the whole C-terminal region (Fig. 4c). This conformational change is coupled to pore opening.

The coupling between the conformational changes in the C-linker and pore regions may be direct. The N termini of the C-linker domains are on the top of the structure and are located a few amino acids after the S6 transmembrane segments, which are proposed to line the pore and form the gate for these channels^{22,23}. The structures of the KcsA and MthK potassium channels present hypothetical conformations for the S6 segments of HCN channels in the closed and open states, respectively^{7,24}. Indeed, the HCN2 structure docks well with the KcsA and MthK structures (see Supplementary Information). The distance of the Asp 443 α -carbon from the central pore axis in HCN2 is 16 Å, which is intermediate between the corresponding distances in KcsA (9 Å) and MthK (~19 Å). Our HCN2 structure is the cyclic nucleotide-modulated conformation of the CNBD and C-linker, and may represent the

open configuration of the channel.

From our studies, a mechanism for the allosteric regulation of HCN channels by cyclic nucleotides is beginning to emerge (Fig. 4c). The initial binding of cAMP is largely in the β -roll and is mediated by electrostatic and hydrogen-bonding interactions with the phosphoribose moiety and by hydrophobic interactions with the adenine ring. cAMP binding produces a conformational change in the CNBD involving the C-helix, as suggested by the finding that C-helix deletions disrupt the allosteric regulation of HCN2 (ref. 8). This conformational change in the CNBD is then coupled to a conformational change in the C-linker, perhaps through the A- and B-helices, that alters the subunit interface between the A'- and B'-helices and C'- and D'-helices of the neighbouring subunit. Finally, a change in quaternary structure and movement of the N-terminal end of the A'-helix, presumably outward, removes the inhibition of the C-terminal region and promotes opening of the ion channel.

Insights gleaned from studies of the HCN2 channel are relevant to other ion channels that are related in amino acid sequence and function (Fig. 1b). In particular, the CNG channels also show a stabilization of the open state on cyclic nucleotide binding⁴. For CNGA1 channels, this stabilization is 8–9 kcal mol⁻¹ with cGMP and results, in large part, from an interaction between the C-helix and the bound cGMP¹². Unexpectedly, however, residues in the A'-helix of CNG channels that show intersubunit Ni²⁺ coordination are 20–30 Å apart in the HCN2 structure²⁵. This may reflect a difference in orientation of the A'-helix in the structure of HCN2 or a state-dependent movement of this helix during gating. The

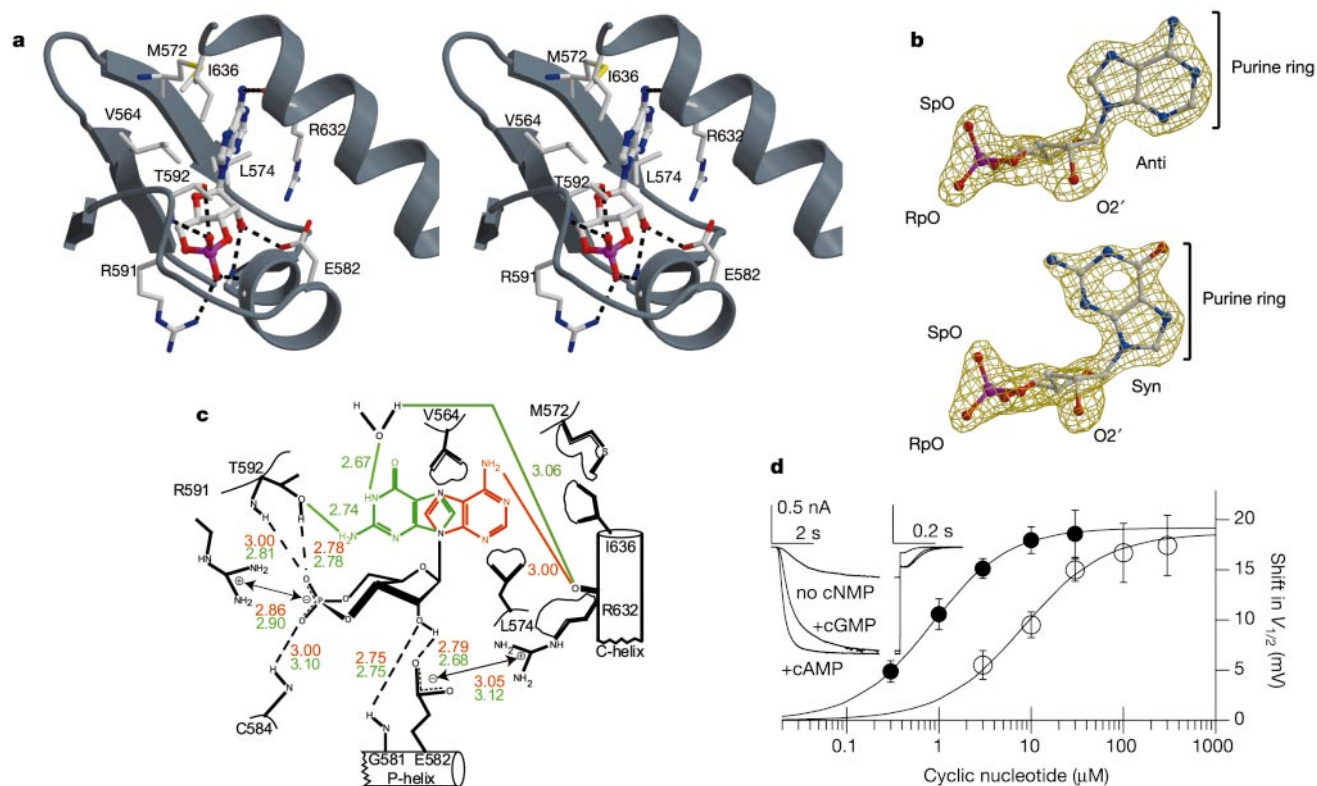


Figure 3 cAMP and cGMP bind to the same site with different stereochemistry.

a, Stereoview of the cAMP-binding site, showing hydrogen bonds and salt bridges (broken lines). **b**, Electron density for cAMP (top) and cGMP (bottom), contoured at 3.3 σ and 5.0 σ , respectively. **c**, Protein–cAMP and protein–cGMP interactions. Broken lines represent hydrogen bonds and double arrows define salt bridges. Green lines delineate hydrogen

bonds formed only with cGMP, red lines show hydrogen bonds formed only with cAMP. Enclosed areas define non-polar interactions. **d**, Shift in $V_{1/2}$ versus concentration of cAMP (filled circles) or cGMP (open circles) for the HCN2 Δ 645 construct. Inset, currents elicited by stepping from –40 mV to –140 mV with or without 10 μ M agonist. Tail currents on return to –40 mV are shown after the axis break.

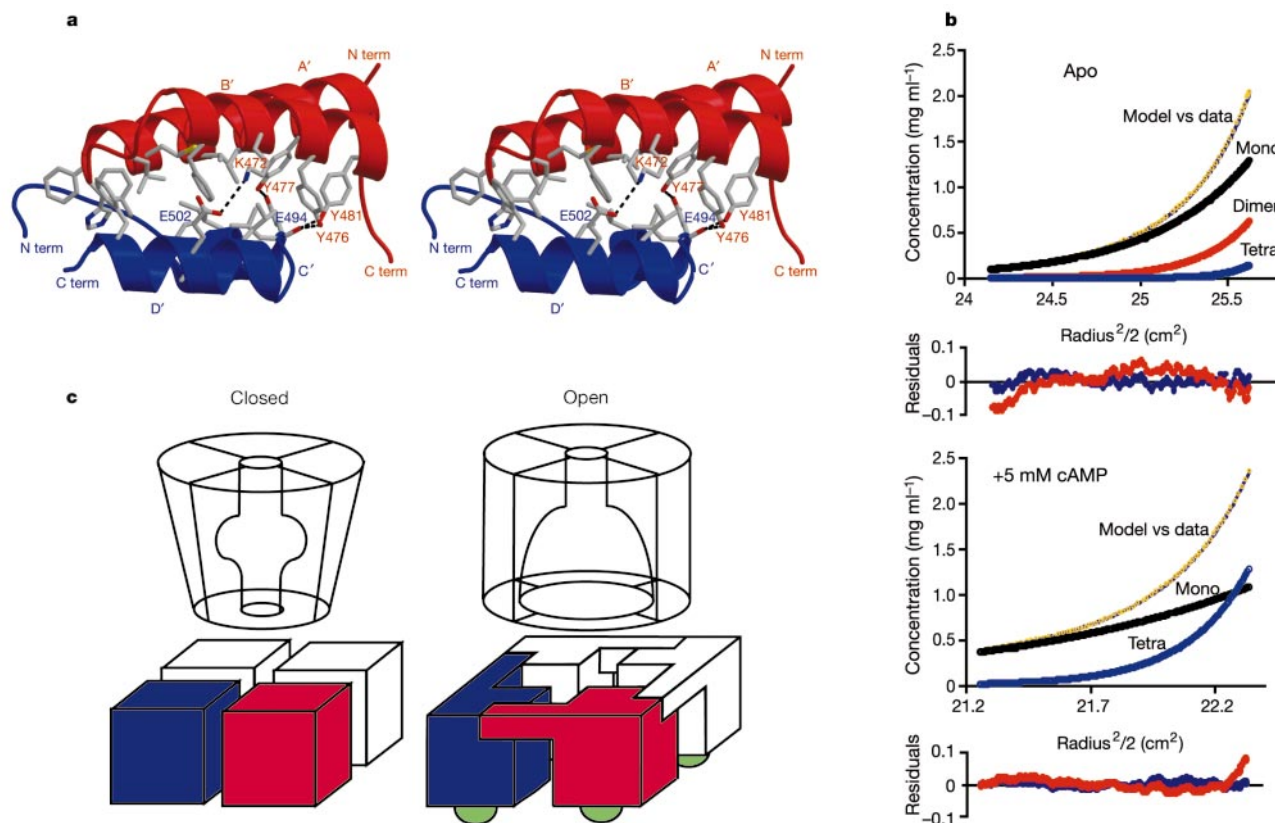


Figure 4 C-linker interactions and cyclic nucleotide-dependent tetramer formation. **a**, Stereoview of intersubunit C-linker contacts. Helices A' and B' of one protomer are in red, and helices C' and D' of an adjacent protomer are in blue. **b**, HCN2I sedimentation equilibrium data in the absence (top) and presence (bottom) of cAMP. The measurements and the models used to fit the experimental data are shown as open yellow circles and thin

black lines, respectively. The thick black, red and blue lines represent, respectively, the proportion of monomer, dimer and tetramer calculated from the models. The residuals are shown for a monomer–dimer (red) or a monomer–dimer–tetramer (blue) model (top), and for a monomer–dimer (red) or a monomer–tetramer (blue) model (bottom). **c**, Cartoon of an unliganded closed channel (left) and a liganded open channel (right).

eag-related families of channels (eag, erg, elk) as well as KAT1 and AKT1 channels, although related in amino acid sequence to HCN channels, lack many of the key residues that interact with the cyclic nucleotides (Fig. 1b) and show weak or no cyclic nucleotide-dependent modulation^{5,6}. These observations suggest the intriguing possibility that CNBD-like modules may couple the binding of small molecules other than cyclic nucleotide to the modulation of ion channel activity. □

Methods

Protein preparation

The DNA fragments encoding residues 443–640 (HCN2I) and 443–645 (HCN2J) of the mouse HCN2 ion channel were cloned into pETGQ. We grew transformed *Escherichia coli* BL-21(DE3) cells to an absorbance of ~0.6 at 600 nm at 37 °C in Luria broth. The cells were then cooled to about 20 °C and expression was induced with 1 mM isopropyl-1-thio-β-D-galactopyranoside. After 5 h, the cells were collected by centrifugation and resuspended in ice-cold lysis buffer (30 mM HEPES, 150 mM NaCl, 10% glycerol, 5 μM cAMP, 1 mM β-mercaptoethanol, 0.1 mM phenylmethylsulphonyl chloride and 2.5 μg ml⁻¹ DNase; pH 7.5).

Working at 0–4 °C, we lysed the cells and cleared the lysate by centrifugation. Imidazole (10 mM) was added to the supernatant, the constructs were purified by Ni²⁺-NTA chromatography, and the octahistidine tag was removed by thrombin cleavage. Subsequently, the thrombin reaction was diluted tenfold with buffer containing 20 mM HEPES, 1 mM dithiothreitol and 5 μM cAMP (pH 7). The solution was loaded onto a S-sepharose ion exchange column, and the protein was eluted with a linear NaCl gradient as two well-separated peaks. The protein in the second peak was generally used for crystallization, although material from both peaks formed isomorphous crystals. We pooled fractions and dialysed them against 20 mM HEPES, 150 mM NaCl, 1 mM dithiothreitol and 5 μM cAMP (pH 7). After dialysis, the protein solution was supplemented with 5 mM cAMP or 5 mM cGMP and concentrated to 5–7 mg ml⁻¹. Selenomethionine derivatives were generated as described²⁶.

Sedimentation equilibrium

Before centrifugation experiments, the protein was purified by size-exclusion chromatography and concentrated to 4.5 mg ml⁻¹. The concentrated protein was separated into two pools and dialysed against buffer containing 20 mM HEPES, 150 mM NaCl and 1 mM dithiothreitol (pH 7) either with 5 mM cAMP or without nucleotide (apo state). Sedimentation equilibrium runs were conducted in a XL-I analytical ultracentrifuge (Beckman/Coulter) with the Rayleigh interference detection system. The protein samples, with and without cAMP, were loaded at concentrations of 4.5, 2.3 and 1.2 mg ml⁻¹ into six-sector side-filling cells. Samples were centrifuged at 12,000, 17,000 and 25,000 r.p.m. in an An50Ti rotor at 4 °C, with scans taken at 1-h intervals until the samples reached equilibrium. Figure 4b shows concentration versus radius plots for the apo state at 4.5 mg ml⁻¹ and 25,000 r.p.m. and for cAMP-bound form at 2.3 mg ml⁻¹ and 17,000 r.p.m. Water blanks were subtracted with WinReedit and data analysis was done with WinNONLIN (<http://spin6.mcb.uconn.edu/winnonlin/winnonlin/html>) and SEDTERP (<http://www.jphilo.mailway.com/default/htm>).

Crystallization and structure determination

We crystallized HCN2J–cGMP, HCN2J–cAMP and HCN2I–cAMP by the hanging drop method at 4 °C. In general, 1 μl of protein solution was mixed with 1 μl of a reservoir solution of 200 mM NaCl, 100 mM citrate (pH 4.6) and 15% PEG 400. Crystals were cryoprotected in reservoir solution supplemented with 20% glycerol and cAMP or cGMP, and flash frozen in liquid nitrogen. Diffraction data from the HCN2I–cAMP crystal were collected at 110 K at MacCHESS, and data sets for crystals of HCN2J–cAMP and HCN2J–cGMP were collected at beamline X4A at National Synchrotron Light Source. Integration, scaling and merging of the diffraction data were done with the HKL suite of programs (<http://www.hkl-xray.com>). A three-wavelength MAD data set was collected on a selenomethionine HCN2I–cAMP crystal. Sixteen out of eighteen selenium sites and initial phases were determined by SOLVE (<http://www.solve.lanl.gov>), followed by density modification with the program DM²⁷. Structures of HCN2J–cAMP and HCN2J–cGMP were determined by molecular replacement using the program AmoRe²⁷ and the HCN2I–cAMP structure as a search probe.

Model building was done with the program O²⁸ using $F_o - F_c$ 'omit' maps and refinement was done with CNS²⁹. Elements of secondary structure were defined by PROCHECK (<http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html>) and the

electrostatic surface calculations were done with GRASP³⁰. The electron density maps shown in Fig. 3b are $F_o - F_c$ omit maps in which the cyclic nucleotide was omitted from the structure factor calculation. Parameters used to define the stereochemistry of cAMP and cGMP were derived from the corresponding small-molecule crystal structures obtained from the Cambridge Structure Database.

Patch-clamp recording of truncated HCN channels

Construction of the complementary DNA for the truncated HCN2 channel (amino acids 1–645; HCN2 Δ 645), expression in *Xenopus* oocytes and voltage-clamp recordings from inside-out patches were done as described¹⁵. Both pipette and bath solutions contained 107 mM KCl, 5 mM NaCl, 10 mM HEPES, 1 mM MgCl₂ and 1 mM EGTA (pH 7.3). cGMP or cAMP was included in the bath solution and applied by gravity perfusion. Currents were elicited by applying 5-s pulses to each of a range of hyperpolarized voltages (in 10-mV steps) and measuring tail currents on return to the holding potential of -40 mV. We corrected tail current amplitudes by subtracting the steady-state holding current and fitted them with the Boltzmann equation, $I = I_{\max}/\{1 + \exp[(V - V_{1/2})/s]\}$, where V is the pulse voltage, $I_{\max}$ is the maximal tail amplitude, $V_{1/2}$ is the activation midpoint voltage, and s is the reciprocal slope. $V_{1/2}$ measurements were taken more than 8 min after patch excision, to allow spontaneous drift in $V_{1/2}$ to be completed, as confirmed by duplicate measurements.

To determine $\Delta V_{1/2}$ (the shift in $V_{1/2}$ induced by cyclic nucleotide agonist), the average of $V_{1/2}$ measurements taken before agonist application and after agonist washout was subtracted from the $V_{1/2}$ measured in the presence of agonist; the mean $\pm$ s.d. values are plotted. Pooled $\Delta V_{1/2}$ data were fitted with the Hill equation, $\Delta V_{1/2} = \Delta V_{1/2,\max}/(1 + K_{1/2}/[A]^h)$, where $[A]$ is the agonist concentration, $\Delta V_{1/2,\max}$ is the maximal shift in $V_{1/2}$, $K_{1/2}$ is the agonist concentration eliciting half-maximal shift and h is the Hill coefficient¹¹.

Received 8 May; accepted 17 July 2003; doi:10.1038/nature01922.

- Robinson, R. B. & Siegelbaum, S. A. Hyperpolarization-activated cation currents: from molecules to physiological function. *Annu. Rev. Physiol.* **65**, 453–480 (2003).
- Biel, M., Schneider, A. & Wahl, C. Cardiac HCN channels: structure, function, and modulation. *Trends Cardiovasc. Med.* **12**, 206–212 (2002).
- DiFrancesco, D. & Tortora, P. Direct activation of cardiac pacemaker channels by intracellular cyclic AMP. *Nature* **351**, 145–147 (1991).
- Zagotta, W. N. & Siegelbaum, S. A. Structure and function of cyclic nucleotide-gated channels. *Annu. Rev. Neurosci.* **19**, 235–263 (1996).
- Bruggemann, A., Pardo, L. A., Stühmer, W. & Pongs, O. Ether-a-go-go encodes a voltage-gated channel permeable to K⁺ and Ca²⁺ and modulated by cAMP. *Nature* **365**, 445–448 (1993).
- Hoshi, T. Regulation of voltage dependence of the KAT1 channel by intracellular factors. *J. Gen. Physiol.* **105**, 309–328 (1995).
- Doyle, D. A. *et al.* The structure of the potassium channel: molecular basis of K⁺ conduction and selectivity. *Science* **280**, 69–77 (1998).
- Wainger, B. J., DeGennaro, M., Santoro, B., Siegelbaum, S. A. & Tibbs, G. R. Molecular mechanism of cAMP modulation of HCN pacemaker channels. *Nature* **411**, 805–810 (2001).
- Weber, I. T. & Steitz, T. A. Structure of a complex of catabolite gene activator protein and cyclic AMP refined at 2.5 Å resolution. *J. Mol. Biol.* **198**, 311–326 (1987).
- Su, Y. *et al.* Regulatory subunit of protein kinase A: structure of deletion mutant with cAMP binding domains. *Science* **269**, 807–813 (1995).
- Chen, S., Wang, J. & Siegelbaum, S. A. Properties of hyperpolarization-activated pacemaker current defined by coassembly of HCN1 and HCN2 subunits and basal modulation by cyclic nucleotide. *J. Gen. Physiol.* **117**, 491–504 (2001).
- Varnum, M. D., Black, K. D. & Zagotta, W. N. Molecular mechanism for ligand discrimination of cyclic nucleotide-gated channels. *Neuron* **15**, 619–625 (1995).
- Matulef, K., Flynn, G. E. & Zagotta, W. N. Molecular rearrangements in the ligand-binding domain of cyclic nucleotide-gated channels. *Neuron* **24**, 443–452 (1999).

- Ludwig, A., Zong, X., Jeglitsch, M., Hofmann, F. & Biel, M. A family of hyperpolarization-activated mammalian cation channels. *Nature* **393**, 587–591 (1998).
- Wang, J., Chen, S. & Siegelbaum, S. A. Regulation of hyperpolarization-activated HCN channel gating and cAMP modulation due to interactions of COOH terminus and core transmembrane regions. *J. Gen. Physiol.* **118**, 237–250 (2001).
- Weber, I. T., Shabb, J. B. & Corbin, J. D. Predicted structures of the cGMP binding domains of the cGMP-dependent protein kinase: a key alanine/threonine difference in evolutionary divergence of cAMP and cGMP binding sites. *Biochemistry* **28**, 6122–6127 (1989).
- Kumar, V. D. & Weber, I. T. Molecular model of the cyclic GMP-binding domain of the cyclic GMP-gated ion channel. *Biochemistry* **31**, 4643–4649 (1992).
- Altenhofen, W. *et al.* Control of ligand specificity in cyclic nucleotide-gated channels from rod photoreceptors and olfactory epithelium. *Proc. Natl Acad. Sci. USA* **88**, 9868–9872 (1991).
- Gauss, R., Seifert, R. & Kaupp, U. B. Molecular identification of a hyperpolarization-activated channel in sea urchin sperm. *Nature* **393**, 583–587 (1998).
- Liu, D. T., Tibbs, G. R. & Siegelbaum, S. A. Subunit stoichiometry of cyclic nucleotide-gated channels and effects of subunit order on channel function. *Neuron* **16**, 983–990 (1996).
- Ulen, C. & Siegelbaum, S. A. Tetramer constructs reveal stoichiometry of cAMP modulation of hyperpolarization-activated HCN2 channels. *Biophys. J.* **84**, 137a (2003).
- Rothberg, B. S., Shin, K. S., Phale, P. S. & Yellen, G. Voltage-controlled gating at the intracellular entrance to a hyperpolarization-activated cation channel. *J. Gen. Physiol.* **119**, 83–91 (2002).
- Shin, K. S., Rothberg, B. S. & Yellen, G. Blocker state dependence and trapping in hyperpolarization-activated cation channels: evidence for an intracellular activation gate. *J. Gen. Physiol.* **117**, 91–101 (2001).
- Jiang, Y. *et al.* Crystal structure and mechanism of a calcium-gated potassium channel. *Nature* **417**, 515–522 (2002).
- Johnson, J. P. Jr & Zagotta, W. N. Rotational movement during cyclic nucleotide-gated channel opening. *Nature* **412**, 917–921 (2001).
- Guerrero, S. A., Hecht, H. J., Hofmann, B., Biehl, H. & Singh, M. Production of selenomethionine-labelled proteins using simplified culture conditions and generally applicable host/vector systems. *Appl. Microbiol. Biotechnol.* **56**, 718–723 (2001).
- Collaborative Computational Project, N. The CCP4 suite: programs for X-ray crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
- Jones, T. & Kjeldgaard, M. Electron-density map interpretation. *Methods Enzymol.* **277**, 173–208 (1997).
- Brunger, A. T. *et al.* Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
- Nicholls, A., Sharp, K. A. & Honig, B. Protein folding and association: Insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11**, 281–296 (1991).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank J. Lidestri for support of the X-ray facility at Columbia University; M. A. Gawinowicz for assistance with protein analysis by mass spectrometry; and S. Siegelbaum for comments and the HCN2 channel cDNA. X-ray diffraction data sets were collected at the MacCHESS synchrotron facility and at NSLS, and we thank the beamline personnel for their assistance. N.O. and R.O. were supported in part by a NIH training grant in Molecular Biophysics. Funding for the analytical ultracentrifuge was provided by the NIH. W.N.Z. and E.G. are investigators with the Howard Hughes Medical Institute.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and request for materials should be addressed to E.G. (jeg52@columbia.edu) or W.N.Z. (zagotta@u.washington.edu). The atomic coordinates for the HCN2I-cAMP, HCN2J-cAMP and HCN2J-cGMP structures described here have been deposited in the Protein Data Bank under the accession codes 1Q43, 1Q50 and 1Q3E, respectively.

Patents, programs and proteins

Personalized search, personalized plasmids.

Science IP

Chemical Abstracts Service

www.sciencelP.org

An on-demand search service

Science IP is a personalized search service for information professionals, attorneys and others involved in patenting and research and development. Information obtained from this source is suitable for supporting decisions on prior art/patentability for the prosecution of new patents, freedom-to-practice/operate for new and existing technologies, and patent validity for infringement litigation. The search draws upon CAS scientists with knowledge of specialized areas, including nucleic and amino acid sequences, polymers, molecular biology and genetics and pharmaceuticals.

DNASIS MAX 2.0

Hitachi Software Engineering

www.hitachisoft-bio.com

Sequence analysis software upgrade

The DNASIS MAX 2.0 software package for Windows includes a wide range of standard sequencing functions. It features automated, customizable sequence analysis, allowing

users to create and save workflows for specific requirements. Enhancements to this version include Contig Manager, a DNA sequence assembly tool for sequencing projects, and Plasmid Map Editor, which can be used to construct the plasmid map supported by various objects such as lines, arrows, labels or texts.

Knockout RNAi System

BD Biosciences/Clontech www.clontech.com

Multiple delivery options for gene silencing

The Knockout RNAi system from BD Biosciences Clontech provides retroviral and adenoviral delivery formats that generate siRNA molecules for effective gene silencing in many cell types. The RNAi-ready pSIREN vectors are ligation-ready plasmids that allow users to choose plasmid or viral delivery. The system overcomes difficulties associated with plasmid-based vectors — namely, the limited ability to deliver siRNA to various cell types. To prepare a construct, users select target sequences within the coding region of the gene of interest and synthesize a double-stranded hairpin oligo for each target sequence. The oligo is then ligated into a pSIREN vector to express the siRNA.

p3 microplate

Porvair Sciences www.porvair-sciences.com

Double feature

The p3 microplate has a dual filter design to prevent sample breakthrough and waste associated with leakages from traditional protein precipitation microplates. Hydrophobic/oleophobic filters retain the entire acetonitrile/sample precipitation mix, allowing complete mixing without loss or leakage. A pre-filter retains large flocculent particles, eliminating 'plugging' problems, while the final filter traps residual protein particles. The large pore size improves flow rate. The p3 processes 96 samples per batch and will integrate with automated systems.

Plasmid Prep Service

Qiagen www.qiagen.com/plasmidprepservice

Prompt preps

Qiagen's Plasmid Prep Service provides ultrapure plasmid DNA in milligram to gram amounts for research and pre-clinical applications. The plasmid DNA is suitable for gene therapy or genetic vaccination projects and for transfection into sensitive eukaryotic cell lines. The company's plasmid purification specialists can help with individual needs, and a range of extra options are available, including low- and ultralow-endotoxin grades, bio-burden, host protein and endotoxin assays and restriction analysis. A certificate of analysis is delivered with each plasmid DNA prep. Preps can be ordered online for rapid turnaround.

These notes are compiled in the Nature office from information provided by the manufacturers.

Programmed for success

Advanced software and services aimed at the non-bioinformatician are making it easier to ride the tidal-wave of genomics and proteomics data. Steve Buckingham reports.

Feel that you're missing out on genomics? Even if you don't think of yourself as 'bioinformatics-enabled', your desktop PC, coupled to the Internet with its ever-expanding bioinformatics resources, will let you tap in to the data deluge. All you need to join the bioinformatics revolution is software that doesn't require you to have a deep working knowledge of bioinformatics and genetics in order to get results. But is it out there? According to software companies and service providers, the answer is a resounding 'yes'.

Finding genes lurking in raw sequence data, determining the proteins that they encode, and managing the flow of data has powered a large bioinformatics industry. Indeed, the software and services offered by big companies such as LION Bioscience in Heidelberg, Germany, coupled with specialist hardware solutions from IBM, Microsoft and Sun, have been indispensable in handling the flow of data coming out of genome-sequencing projects. But it is not just a problem of scale. As well as transforming biological knowledge, the rapid emergence of such a volume of raw data is also changing the way in which researchers

frame the questions that shape research projects and the design of experiments.

But is this empowering knowledge getting to all of the researchers who would benefit from it? Biologists want data in a form that is meaningful to their own research, but they don't always have the time or resources to master the art of bioinformatics or to hire someone who has. So how can bench scientists in a small or medium-sized company or university department cash in on all of this raw information?

Help where it's needed

The need now is for software and services that can broker this information and pass it on to the non-expert in an intuitive way. "The challenge is that in-depth expertise has to be packaged into an environment that appeals to the average biologist, is easy to understand and easy to use," says Klaus May, director of sales and marketing for Genomatix Software in Munich, Germany.

A growing bioinformatics market is trying to bridge this gap. There are encouraging signs of a trend towards software with intuitive interfaces coupled with ever-increasing analytical power. David Speechly,

vice-president of the life-sciences company Applera in Norwalk, Connecticut, compares the situation to the early days of motoring. "The first cars were driven by engineers and mechanics," he says. "Later, everyone was able to drive them." Already, Speechly notes, new tools are opening up the human genome across the life sciences.

Take, for example, someone working on cardiovascular disease. They can now run a computer search of the more than 5 million known human single-nucleotide polymorphisms (SNPs, pronounced 'snips', see *Nature* **422**, 917–923; 2003) for those associated with the condition. They can then test their control groups for these SNPs to screen out those people who might have unidentified disease, thus removing one source of uncertainty from their experiment. Applera and some other companies allow academic researchers free access to their SNP database for this kind of search, and offer a catalogue of SNP-testing kits.

The bioinformatics industry itself has had to meet some tough challenges. Ron Ranauro, general manager of the Paris, France-based company Gene-IT, sees a pressing need for software firms to produce

FINDING YOUR WAY

Most genome databases now feature more than just basic sequence information; they encompass a wide range of annotated information on factors such as disease states, protein structure and polymorphisms. Yet many researchers remain unaware of the potential of these resources. A much-cited survey commissioned two years ago by the London-based Wellcome Trust revealed that only about half of the biomedical researchers

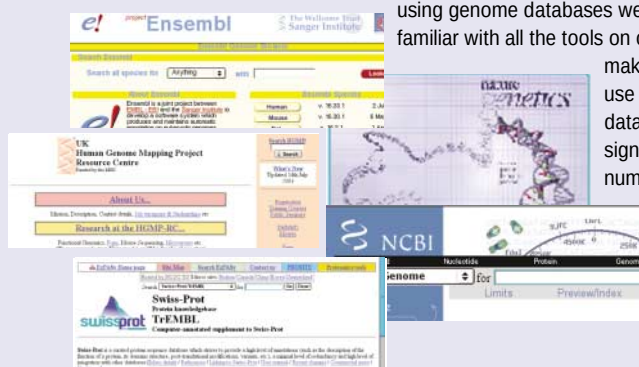
using genome databases were familiar with all the tools on offer to make full use of the data. A significant number

didn't even know of the existence of the European Bioinformatics Institute's public Ensembl database and website, which collates vast amounts of diverse information on the human genome. The findings prompted the trust to launch an educational initiative in 2001 to draw attention to these databases.

So where do you turn to make sense of the number of databases available and the array of tools they offer? A good starting point is the annual database special issue of *Nucleic Acids Research* (**31**; 1–516; 2003). The online version of this issue contains an index of major databases along with a brief summary of what each offers, information that is compiled by Andreas Baxevas of the genome-technology branch at the US National Human Genome Research Institute.

Newcomers are often overwhelmed by the tools on offer and the knowledge of genetics that is presupposed on the part of the user. The tools tend to assume an understanding of terms and principles that are unfamiliar to, say, physiologists and pharmacologists, and without this knowledge the users cannot feel confidence in the results that they get. A useful guide for the perplexed is *A User's Guide to the Human Genome*, a web special published by *Nature Genetics* (www.nature.com/ng). This is designed for the genome neophyte, and guides the user through worked examples to give them confidence in understanding concepts and strategies that can then be used to empower their own research.

S.B.

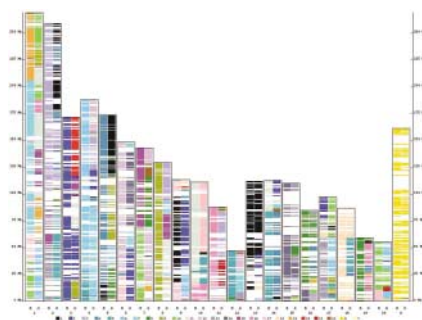


Genome gateways: but how to get into the garden?

programs that are more adaptable to different users. "It is definitely not a case of one-size-fits-all," he says. "The way applications have to be developed now is to work from the consumers' needs backward." Gene-IT's new product, GenomeQuest, to be released this month, aims to provide an intranet-based sequence-search solution that makes it easy for scientists to get a complete picture of functional annotation from the entire sequence world and coordinate sequence information across diverse research teams.

There are other challenges. "The main problem, as I see it, is the pace of discovery," says Bill Ladd, senior director of analytic applications for software developer Spotfire in Somerville, Massachusetts. "Bioinformatics exists to support very dynamic, and therefore very different, claims on software development. The technology is improving all the time, and so the analysis changes. In the past, when something new came along you would have probably a few months to gather the requirements and another few months or so to roll out the new software. Now that cycle has sped up to a matter of weeks, if not days."

And the pace is only going to accelerate further. The way in which software interacts with the user has undergone a sea-change in the past few years, largely in response to the need to deal with large data sets. Ladd believes that there has been a migration towards the use of web interfaces because they are easy to manage and develop. But this has a cost. "It means that we are doing more browsing than analysis," he says. "There are databases like Ensembl, for example, that effectively collate data from



Three-way synteny from Softberry.

different sources, but difficulties arise when you try to integrate even this collated database data with experimental data. Once I have a list of genes out of Ensembl, what happens when I ask whether other genes in my data have the same characteristics?"

New lamps for old

Fortunately help is at hand for the non-expert. Many of the new bioinformatics products aim to do essential tasks, such as BLAST queries, which find matching sequences in databases, and protein alignments, more efficiently and in a more user-friendly way than previous systems. New algorithms allow searches of large genomes to be done at unprecedented speeds without the loss in sensitivity that results in missed alignments. This means that it is possible to do real-time interactive searches against whole genomes and genomic data.

One such package is PatternHunter from Bioinformatics Solutions in Waterloo,

Canada, which introduces the concept of 'spaced seed' to accelerate homology searches, and claims to be some 100 times faster than traditional BLAST. Whereas BLAST looks for regions where 11 consecutive residues match, PatternHunter looks for any 11 matches over, for example, an 18-residue segment, making the search more sensitive and, surprisingly, faster. Using a multiple-seed approach gives PatternHunter the sensitivity of the Smith-Waterman algorithm, but up to thousands of times faster. It runs on Sun Microsystems' Java Virtual Machine and also boasts conservative memory usage and a guarantee not to miss any alignment. The program was used by the Mouse Genome Sequencing Consortium to compare the mouse and human genomes (*Nature* **420**, 520–562; 2002), and is also used by companies such as Celera Genomics in Rockville, Maryland, and deCODE Genetics in Reykjavik.

Fast searching is also a feature of Genome Explorer from Softberry in Mount Kisco, New York, which uses the FMAP algorithm. This is a very fast algorithm developed by Softberry to map query sequence to large genomes. It keeps the oligonucleotide vocabulary of the entire genome in computer memory to speed up the searches. Softberry claims that the program can search the entire human genome for a sequence of interest in under two seconds. As well as offering simple pattern searches, retrieval of nucleotide and amino-acid sequences, Genome Explorer includes access to expression data on genes and the annotation of the draft of the human genome.

A number of desktop programs make it

SEEING IS BELIEVING

Science often works by data mining — finding correlations within and between sets of data. Dividing these sets into subsets and testing out various scenarios are key steps in this process. But many data sets generated today are extremely large and complex, sometimes involving several dimensions and varying levels of subdivision.

And this is not only a concern for large pharmaceutical companies — anyone using microarrays has the same problem.

Many new bioinformatics products address this problem by organizing data in a dynamic visual context. "People are visually oriented. They are more productive when data are presented visually rather than textually," says Ron Ranauro, general manager at Paris-based Gene-IT. Data visualization aims to allow scientists to get an intuitive grasp of data structure and to spot potentially interesting trends. For example, the well known SigmaPlot package made by SPSS in Chicago, Illinois, is probably used as much for trends analysis and scenario testing as it is for the preparation of graphs for publication.

Spotfire in Somerville, Massachusetts, prides itself on the ease of use of its data visualization and decision-making software, such as DecisionSite. The functional genomics version of this program allows users to visualize genomics data and spot trends and correlations. It accepts data from a

wide variety of different databases, addressing the old problem in bioinformatics that relevant data are dispersed across different locations and are often in widely divergent, and potentially incompatible, formats.

Data visualization tries to shorten the path to the 'eureka!' moment, where the researcher has intuitively grasped what the data are saying. But intuition must be backed by rigorous analysis. Programs are often packaged with a number of powerful analytical tools including similarity searches, replicate summarization and coincidence testing. DecisionSite, for example, comes with preconfigured guides to assist in common genomic analyses such as gene finding, generating "heat maps" — a type of visualization where data is colour-coded to enable an overview of large amounts of data at once.

S.B.



Ron Ranauro: spotting trends.

easy to do protein and nucleic-acid alignments, as well as to design primers, to search motifs and to perform multiple sequence analysis. One example is MacVector, for the Apple Macintosh and the Windows version DS Gene, produced by Accelrys in San Diego, California. MacVector has been available for many years and is continually being improved. The reasonably priced Jellyfish from LabVelocity in San Francisco, California, is also available for both Mac and PC. Like MacVector, Jellyfish will generate primers, oligonucleotides, cloning constructs and restriction maps, as well as perform BLAST searches and sequence alignments. When sequence data are downloaded, so are the annotations. Another option is the Visual Cloning 3 package from Redasoft in Bradford, Ontario, which the company claims simplifies sequence analysis, sequence editing and presentation, as well as offering access to online tools through its integrated web-browser interface and an array of plain-language 'wizards' to guide the user through the process.

Proteomics has benefited from software that not only allows rapid analysis of two-dimensional electrophoresis gels, but also draws together other functions, such as data mining, into one easy-to-use package. Progenesis from Nonlinear Dynamics in Newcastle upon Tyne, UK, comes in three modules. The first allows data annotation, the second does the image analysis and data logging, and the third is the data-mining component in which, for example, the data from different gels can be compared. Users can generate pick-lists of interesting spots



Making sense of 2D gels.

from the data-mining results and send them back to the image-analysis module to drive spot-picking from the original gel.

Paddy Lavery, Nonlinear's bioinformatics marketing manager, says that a future version of the software will allow users to import raw mass-spectrometry traces or peptide map lists into the program and to search against internal or external peptide sequence databases. The user will be able to store the search results and link them back to the gel and the sample ID of any spot in the pick list.

Making predictions

As the number of sequences of known function increase, predictions of genes, protein structure and protein function from sequence information are getting more accurate. There is a trend towards collections of integrated, coordinated suites of gene-prediction programs, many of which can be tried out on the web. Softberry, for example, offers a number of gene-prediction programs that can be accessed over the web, including FGENESH,

which the manufacturers claim is fast, sensitive and accurate. The suite also includes FGENESH_C, which searches for similar cDNAs, and FGENESH+, which will find similar proteins. The fully automated FGENESH++C will automatically annotate any genome (other than human) to a standard claimed to be indistinguishable from manual annotation, using a battery of complementary techniques.

The accuracy of programs that predict protein structure from sequence has improved over the past few years and these are gradually becoming more widely used. A number of easy-to-use academic and commercial protein-structure prediction programs are available. PROSPECT Pro from Bioinformatics Solutions uses the 'threading' method, which threads the query sequence onto all known protein folds from the Protein Data Bank to find the one that gives the best-fitting structure. The program builds on this established strategy by allowing the user to feed in experimental data, such as constraints on the threading, and it checks its own results using a neural network.

Predicting the localization of a protein within the cell is also a help in identifying the function of a gene product. Softberry's ProtComp illustrates the trend in bioinformatics software to integrate diverse computational approaches. The program uses clues in a protein's sequence to guess at where it is localized within a cell, mostly by using neural networks to check sequence elements for tell-tale localization-specific motifs.

Another important issue is monitoring and controlling the flow of data as they are

GENOMIC MERGERS

According to Celera Genomics in Rockville, Maryland, the next step after the sequencing of the human genome will be 'merging technologies'. Celera, recently forged agreements with Applied Biosystems in Foster City, California, under the umbrella of the Applera Corporation to work towards integrating all aspects of genomics. The companies are now developing an array of predesigned and prevalidated assays for genes identified on completion of the sequencing of the human genome. There are already assays for more than 18,500 human genes known to be expressed, as well as kits for over 125,000 single-nucleotide polymorphisms (SNPs) — their goal is to have 200,000 SNP assays by the end of the year.

"The most common complaint we hear from scientists is: 'but we're not bioinformaticians'," says Ramin Cyrus, a senior director for Celera. Anthony Kerlavage, a senior director at Celera, likens the situation to modern word processors. "They are packed with features, but most of us just use them to write letters," he smiles. The solution? This month, Celera and Applied Biosystems launched the 'myScience' portal. This website allows users to upload data from instruments, databases or laboratory information management systems (LIMS), and store them in personalized workspaces. Users can then analyse their data with an array of tools, guided by predesigned workflows. These workflows act like 'wizards' — guiding the



Which gene? Assays for human genetics are big business.

user through the options of which web-based tools are available. The site itself is free (Applied Biosystems hopes that it will tempt users to purchase the company's assays), and is designed to complement the subscription-based Celera Discovery System, which offers deeper analyses.

S.B.

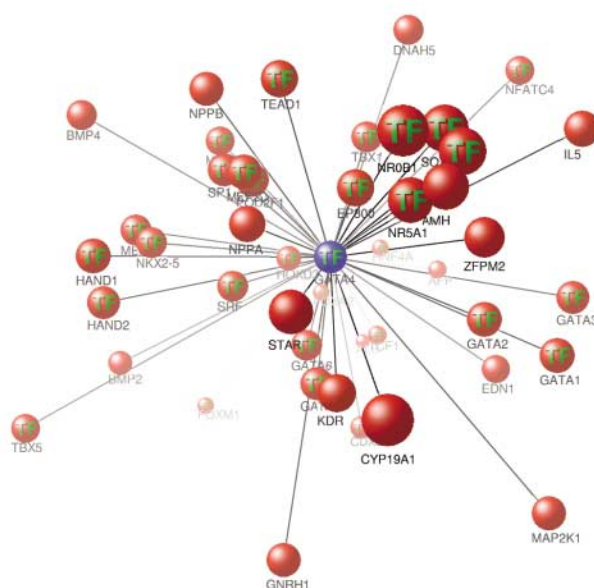
generated and passed through the laboratory, especially as individual experiments can now produce large, complex data sets that need to be interpreted and passed on to the next set of experiments. This is where LIMS come in — laboratory information management systems that not only store data, but will track what reagents are used and where they are bought and stored, as well as following the project's progress. LIMS have been described as electronic lab notebooks — a good LIMS package will catalogue experimental data, which it can capture directly from laboratory instruments, scientists' reports and even purchasing. Another essential feature is traceability — original data must be open to tracking and retrieval throughout the project.

John Helfrich, programme manager for drug discovery and development at NuGenesis Technologies in Westborough, Massachusetts, has been keeping track of how LIMS are evolving. "I am seeing a trend towards the development of 'purpose-built' LIMS that perform a specific subset of data management within specific departments in the biopharmaceutical industry," he says. He points to the Watson LIMS from InnaPhase in Philadelphia, Pennsylvania, which was specifically designed for preclinical bioanalysis in drug discovery.

Another likely trend will see LIMS packages become easier to configure — traditionally a LIMS has been designed in consultation with the client to meet a specific need, but Helfrich thinks that future LIMS vendors will aim at more customizable off-the-shelf products. The latest version of NuGenesis' interoperable SDMS (Scientific Data Management System) was launched this June, with enhanced support for the Macintosh OS X, UNIX and Windows platforms. The system catalogues and captures all of a project's data from its source, thus avoiding the problem of traditional LIMS designed to capture only a narrowly limited data stream or restricted to using treated data. It can be integrated with an existing LIMS or any high-order IT system, or implemented as a stand-alone system for the small or medium-sized lab.

Spending less time in the library

The explosive increase in sequence data is almost matched by the increase in text publications, and keeping up with the published literature can be a full-time task. Some companies are now producing software that allows the user to explore textual data at a level beyond a simple literature



Joining the dots: BiblioSphere from Genomatix searches the literature for co-citations of genes.

search. PubGene, based in Oslo, Norway, offers a program at both free and proprietary levels that identifies potential associations of genes and proteins by finding their co-occurrence in abstracts of published papers or in gene-expression experiments. The publicly available version allows searches for co-occurrences of genes, although the database behind the commercial version is claimed to be more up-to-date and also offers protein searches.

A suite of programs from Genomatix Software takes a similar approach. One powerful member of this suite, which illustrates the emerging emphasis on the visual presentation of complex data, is BiblioSphere. This package brings together genome analysis and the US National Center for Biotechnology Information's PubMed database. Like PubGene, BiblioSphere is a data-mining tool that looks for co-citations of two or more genes of interest in published abstracts. The lowest-level search tags the citation of two genes in the same abstract. The most discerning search looks for the co-citation of the two genes in the same sentence, coordinated by a key word such as 'regulates'. The findings are presented graphically — the gene you searched for is presented visually at the centre of a sphere of related genes. Clicking on the line joining two genes leads you to the citation mentioning the two genes. Clicking on any gene leads you to further data on that gene.

Some programs even claim to be able to track that most complex domain of biological information — the research paper — to find exactly the papers you need. Despite the obvious difficulties faced by a machine attempting to 'understand' natural, human language, some packages claim to do just

that. The LexiQuest Knowledge Management Suite from SPSS in Chicago, Illinois, is based on 'real' linguistics: it can race through unstructured text and produce a graphical map of the main concepts in that text. SPSS claims that the software understands natural language queries and can respond to them intelligently. "Everyone is calling text 'unstructured data', but that's not quite true," says Catherine DeSesa, senior analyst at SPSS. "Text does have structure because language has structure — a very complex structure — and it is the incorporation of the knowledge of this structure that enables LexiQuest Mine to accurately extract and organize multi-word concepts without prior knowledge of the exact terms themselves." Another example is KDE TextSense from InforSense in London, UK, which

contains at its core a free-text mining toolbox and can be used to turn the information into structured data tables.

New information is driving software development, but changes are also needed to the way in which data are presented. Take the database problem — a huge number of databases now exist, all of varying quality and featuring different, often incompatible, formats. Matthew Day, databases editor for London-based online publishers BioMed Central, sees the need for change. "There aren't yet public repositories for all the different sorts of information that biologists are producing. I believe that all data sets should be published as user-friendly online databases that are closely associated with journal articles and are amenable to data mining. Thus the boundaries between journals and databases becomes blurred, and a sea of data sets is created under the umbrella of the peer-review system."

In the genome age, it is easy for small laboratories to feel that they are being left behind. Large biotech companies can now achieve in an afternoon what used to take an average-sized lab months, if not years, to accomplish. But software technologies are getting better at encapsulating expertise into compact programs. These are becoming easier to use and more reliable, and the results are easier to interpret. There is a growing emphasis on automatically drawing on diverse data sources from widely divergent locations. New services and software products are drawing us closer to the day when the expertise of a professional bioinformatician can be downloaded into a desktop computer.

Steve Buckingham is a neurophysiologist at the Department of Molecular Biophysics, University of Oxford. He is also a freelance writer.

Company	Products/activity	Location	URL
Sequence, genome and gene-expression analysis			
Accelrys	GCG Wisconsin Package; MacVector for sequence analysis; Discovery Studio suite for Windows for database mining, genomics and proteomics	San Diego, California	www.accelrys.com
Affibody	Software for genomics data analysis and management	Bromma, Sweden	www.affibody.com
Affymetrix	GeneChip microarray systems and gene-expression analysis tools	Santa Clara, California	www.affymetrix.com
ApoCom Genomics	Desktop bioinformatics tools for gene prediction (GrailEXP, GeneHound) and gene-expression analysis (Excavator)	Knoxville, Tennessee	www.apocom.com
Array Genetics	Online proteomics information database; bioinformatics tools for genomics, proteomics and microarray analysis	Newtown, Connecticut	www.arraygenetics.com
Biocomputing	Data-management systems for genotyping and phenotype data	Espoo, Finland	www.biocomputing.fi
Bioinformatics Solutions	Desktop bioinformatics software: PatternHunter for DNA/protein analysis and PROSPECT for protein-structure prediction	Waterloo, Canada	www.bioinformaticssolutions.com
BioTools	Analysis software for gene and protein sequences and chromatograms	Edmonton, Canada	www.biotools.com
Cognia	Bioinformatics software, including BIOBASE software and databases	New York, New York	www.cognia.com
Curagen	GeneScape portal for genome analysis tools	Branford, Connecticut	www.curagen.com
Digital Gene Technologies	TOGA gene-expression analysis software	La Jolla, California	www.dgt.com
DNASTAR	Lasergene sequence analysis software for Macintosh and Windows; GenVision genomic data visualization tool	Madison, Wisconsin	www.dnastar.com
Entigen	BioNavigator platform for sequence and genome analysis	Sunnyvale, California	www.entigen.com
GATC Biotech	Accelrys, DNASTAR and other bioinformatics software; DNA sequencing	Constance, Germany	www.gatc-biotech.com
GeneBio	Commercial access to SWISS-PROT databases; bioinformatics software	Geneva, Switzerland	www.genebio.com
Gene Codes	Sequencher sequence assembly and analysis software	Ann Arbor, Michigan	www.genecodes.com
Gene-IT	Universal software for sequence database management, comparison, clustering and results analysis	Le Chesnay, France	www.gene-it.com
Genomatrix Software	Bibliosphere for literature searches and data mining; software for genome annotation, analysis and retrieval of promoters, and genome analysis	Munich, Germany	www.genomatrix.de
Genomic Solutions	Genomics and proteomics bioinformatics tools	Ann Arbor, Michigan	www.genomicsolutions.com
Geospiza	Finch servers and tools for sequence assembly, analysis and data management	Seattle, Washington	www.geospiza.com
Hitachi Software Engineering	DNASIS desktop bioinformatics software for DNA sequence assembly and analysis, and analysis of microarray data	Yokohama, Japan	www.hitachi-sk.co.jp/English/index.html
Inpharmatica	Biopendium and CelerEdition Biopendium proteome annotation resources; PharmaCarta chemogenomics informatics platform	London, UK	www.inpharmatica.com
InforMax	Vector bioinformatics software for sequence, genome and microarray data; Vector NTI for Macintosh; LabShare for data storage and management	Bethesda, Maryland	www.informaxinc.com
Iobion Informatics	GeneTraffic microarray data-management and analysis software	La Jolla, California	www.iobion.com
iSenseit	iOmega microarray data analysis and storage software; oligonucleotide computation	Bremen, Germany	www.isenseit.com
LabBook	eLabBook web-enabled electronic notebooks; annotated human genome database and data-mining tools	McClean, Virginia	www.labbook.com
LabVelocity	Jellyfish desktop bioinformatics software; information services	San Francisco, California	www.labvelocity.com
LION Bioscience	Bioinformatics software, database development and management; DiscoveryCenter platform for data integration; contract bioinformatics	Heidelberg, Germany	www.lionbioscience.com
MiraiBio	DNASIS desktop software for DNA sequence assembly and analysis, protein sequence analysis, and analysis of microarray data	Alameda, California	www.miraibio.com
Molecular Biology Insights	Oligonucleotide identification software	Cascade, Colorado	www.oligo.net
Paracel	Software for sequence assembly, analysis and sequence-based genotyping	Pasadena, California	www.paracel.com
Premier Biosoft	Desktop bioinformatics packages for sequence analysis, primer design and two-hybrid protein interactions	Palo Alto, California	www.premierbiosoft.com
PubGene	PubGene public-access and commercial gene databases and analysis software	Oslo, Norway	www.pubgene.com
Redasoft	Visual Cloning 3 software for sequence management and analysis	Toronto, Ontario	www.redasoft.com
Rosetta BioSoftware	Rosetta Resolver and Rosetta Luminator gene-expression data-analysis system	Kirkland, Washington	www.rosettabio.com
Sagitus Solutions	MaxD data warehouse, analysis and visualization software for microarray data	Manchester, UK	www.sagitusolutions.co.uk
science factory	BRENDA enzymology database; überTOOL bioinformatics platform for sequence, expression and structural data	Cologne, Germany	www.science-factory.com
Silicon Genetics	MetaMine, GeNet and GeneSpring microarray analysis software	Redwood City, California	www.sigenetics.com
Softberry	Software for sequence and genome analysis and database searching	Mount Kisco, New York	www.softberry.com
Southwest Parallel Software	Bioinformatics software packages	Albuquerque, New Mexico	www.spsoft.com
Spotfire	DecisionSite analytical and statistical data-management software	Somerville, Massachusetts	www.spotfire.com
SPSS	Clementine statistical and data-mining software; LexiQuest literature text-mining and categorization tools	Chicago, Illinois	www.spss.com
Textco	Gene Construction Kit and Gene Inspector desktop bioinformatics packages and electronic lab notebook	West Lebanon, New Hampshire	www.textco.com
X-MINE	Bioinformatics platform storage and analysis of genomics data	Brisbane, California	www.XMine.com
Databases			
Ardais	BIGR library of biomaterials and information for genomics-based biomedical research	Lexington, Massachusetts	www.ardais.com
Beilstein Information	Chemical databases	San Leandro, California	www.beilstein.com
BIOBASE	TRANSFAC family of databases; analysis tools for gene expression, promoters and signalling pathways; contract bioinformatics	Wolfenbüttel, Germany	www.biobase.de

Company	Products/activity	Location	URL
Biomax Informatics	Annotated human-genome database; customized data management	Martinsried, Germany	www.biomax.de
BioWisdom	Text search and pharmacology and oncology information databases	Cambridge, UK	www.biowisdom.com
Celera Genomics	Celera Discovery System for accessing the Celera annotated genomes databases; bioinformatics services	Rockville, Maryland	www.celera.com
Compugen	GenCarta annotated human genome, transcriptome and proteome database	Tel-Aviv, Israel	www.cgen.com
DECODON	Software for 2D-gel analysis and information storage	Greifswald, Germany	www.decodon.de
GeneLogic	Gene-expression databases and software for drug discovery	Gaithersburg, Maryland	www.GeneLogic.com
Iconix	DrugMatrix databases and software platform for chemogenomics research	Mountain View, California	www.iconixpharm.com
Incyte Genomics	LifeSeq and ZooSeq ESTdatabases; BioKnowledge Library protein-information databases; bioinformatics software	Palo Alto, California	www.incyte.com
Lexicon Genetics	Gene-knockout and gene-function databases and bioinformatics for drug discovery	The Woodlands, Texas	www.lexgen.com
LifeSpan BioSciences	Gene-expression and protein-localization databases and data-mining tools	Seattle, Washington	www.lsbio.com
MDL	Biological and chemical information databases; data-management software	San Leandro, California	www.mdli.com
Cergent	Protein and protein-structure databases; computational proteomics	San Diego, California	www.strubix.com

Computer systems, middleware and laboratory information management systems (LIMS)

Amersham Biosciences	Sierra Laboratory Workflow Systems for microarray, genotyping, sequencing and 2D-MS data	Piscataway, New Jersey	www.amershambiosciences.com
CLONDIAG	PARTISAN microarray LIMS	Jena, Germany	www.clondia.com
geneticXchange	discoveryHub middleware platform for biological data integration	Menlo Park, California	www.geneticxchange.com
HeliXense	Genomics Research Network Architecture (gRNA) software and system infrastructure supporting biological data management	Singapore	www.helixense.com
Hewlett-Packard	Computer systems, networks and data-management systems for the life sciences	Palo Alto, California	www.hp.com/technicalsolutions
IBM	DiscoveryLink platform for database integration; data-management systems	White Plains, New York	www.ibm.com/solutions/lifesciences
InforSense	Kensington Discovery Edition research-management software platform; KDE TextSense	London, UK	www.inforsense.com
InnaPhase	LIMS for drug discovery	Philadelphia, Pennsylvania	www.innaphase.com
Labtronics	Laboratory software and control systems, LIMS interfacing	Guelph, Ontario, Canada	www.labtronics.com
L.I.M.S.	StarLIMS laboratory information management system	South Hollywood, Florida	www.starlims.com
Mitsui Knowledge Industry	LIMS; software for membrane protein secondary-structure prediction, data management and analysis of gene-expression and SNP data	Tokyo, Japan	bio.mki.co.jp
NEC	Computer systems and networks	Tokyo, Japan	www.nec-global.com
NuGenesis	SDMS data management system; scientific data management services	Westborough, Massachusetts	www.nugenesis.com
ProteDyne	LIMS middleware for integration of network-enabled laboratory software	Martinsried, Germany	www.proteDyne.com
Silicon Graphics	SGI servers for high-throughput computing, visualization and data management	San Francisco, California	www.sgi.com
Sun Microsystems	Servers and workstations for high-throughput computing; universal software platforms for networks	Santa Clara, California	www.sun.com
TimeLogic	DeCypher system for accelerated bioinformatics	Crystal Bay, Nevada	www.timelogic.com
TurboWorx	Open computational platforms for biological research data including bioinformatics	Burlington, Massachusetts	www.turbogenomics.com

Services

Aber Genomic Computing	Design of data-mining and predictive modelling software	Aberystwyth, UK	www.abergc.com
AGOWA	Genome analysis; automated sequence annotation bioinformatics services	Berlin, Germany	www.agowa.de
BioInformatics Services	Computational biology; bioinformatics services	Rockville, Maryland	www.bioinformatics-services.com
Chemical Computing Group	Bioinformatics software, services and computer-aided molecular design	Montreal, Quebec, Canada	www.chemcomp.com
Cyberell	Bioinformatics software and services	Helsinki, Finland	www.cyberell.com
ePitope Informatics	Epitope prediction over the web	Durham, UK	www.epitope-informatics.com
GeneData	Bioinformatics systems and services; database development and management	Basel, Switzerland	www.genedata.com
Genometrix	Genotyping, gene expression and bioinformatics services	The Woodlands, Texas	www.genometrix.com
Keygene	DNA fingerprint analysis software; contract genomics and bioinformatics services	Wageningen, The Netherlands	www.keygene.com
SRI International	Contract informatics services	Menlo Park, California	www.sri.com
Tripes	Molecular modelling and virtual-screening software; contract informatics	St Louis, Missouri	www.tripos.com

General

Applied Maths	Gel fingerprint analysis and bioinformatics software; contract bioinformatics	Austin, Texas	www.applied-maths.com
Applied Biosystems	Instruments and assays for genomics, proteomics and molecular biology; myScience website for genome analysis	Foster City California	home.appliedbiosystems.com
Bioalma	Bioinformatics software, software development, consultancy and training	Madrid, Spain	www.almabioinfo.com
Bio-Rad	WorksBase bioinformatics software for proteomics	Hercules, California	www.discover.bio-rad.com
BioSolveIt	Molecular modelling, docking, and protein threading software; services	St Augustin, Germany	www.biosolveit.de
Dalicon	Bioinformatics software for large-scale data management and analysis	Nijmegen, The Netherlands	www.dalicon.com
Elsevier Science	Scirus web-search engine for scientists	Amsterdam, The Netherlands	www.scirus.com
Invitrogen	Kits and reagents for cloning, genomics, genotyping, proteomics, molecular and cell biology	Carlsbad, California	www.invitrogen.com
MegaMetrics	Data-mining software for microarray, proteomics and SNP databases	Wyndmoor, Pennsylvania	www.megametrics.com
Molecular Mining	Data-mining software	Kingston, Ontario, Canada	www.molecularmining.com
Partek	Pattern-recognition and interactive visualization software; consulting services	St Charles, Missouri	www.partek.com

● see advertisement

Contacts

Publisher: Ben Crowe
Editor: Paul Smaglik
Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Global Head of Naturejobs:
Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)

Andy Douglas (4975)

Frank Phelan (4944)

Netherlands/ Italy/ Spain/ Portugal/ Belgium:

Evelina Rubio Hakansson (4973)

Scandinavia: Sille Opstrup (4994)

France/ Switzerland:

Amelie Pequignot (4974)

Natureevents:

Paul Constant (4954)

Production Manager:

Billie Franklin
To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Office

Germany/ Austria:

Patrick Phelan, Odo Wulffen

Tel + 49 89 54 90 57 11/-2

Fax + 49 89 54 90 57 20

e-mail: p.phelan@nature.com

o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,

10th Floor, New York, NY 10010-1707

Tel +1 800 989 7718

Fax +1 800 989 7103

e-mail: naturejobs@natureny.com

US Sales Manager:

Peter Bless

US Advertising Coordinator:

Linda Adam

Japan Head Office, Tokyo

MG Ichigaya Building (5F),

19-1 Harakakamachi,

Shinjuku-ku,

Tokyo 162-0841

Tel +81 3 3267 8751

Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami

e-mail: rasami@naturejpn.com

naturejobs

Fact or fiction?

An advert promoting the services offered by *Naturejobs* has provoked a certain amount of criticism. The ad, which shows fictional postdoc Dr Ramirez 'searching' for a job while enjoying a massage, has prompted real postdoc Maria Sideris to write in. Ramirez seems all too relaxed, she says. More importantly, how can an out-of-work postdoc afford a massage? Postdocs nowadays have student loans to pay off. And even when they secure a fellowship, they often have to live on a stipend lower than the entry-level salaries for graduates.

Sideris, a postdoc at Ohio State University's department of molecular genetics in Columbus, goes on to make some valid points about the plight of young researchers — particularly women. Ramirez would surely be more stressed if she were a mother, Sideris writes, especially a mother without access to affordable childcare, which is a reality for researchers in many universities. Or maybe Ramirez is holding off from having children, as women researchers sometimes feel that becoming a mother can derail the tenure-track train.

In her letter, Sideris also questions the high-tech approach that the ad touts of letting web-based 'job agents' scour databases for suitable positions. Of course, the ad doesn't say job-seeking scientists should rely only on high-tech tools — old-fashioned methods such as networking, tracking research-funding trends, and contacting possible mentors and collaborators still apply.

Although the ad's storyline may be a little over the top, its exaggeration helps to put its point across. But it is important to distinguish between fact and fiction, moderation and hyperbole, even when they sometimes come together. As they do in Sideris' wish for her ad-world counterpart, which nicely echoes *Naturejobs*' wish for its readers: "I wish them all the best in their quest for professional stability and success."

Paul Smaglik

Naturejobs editor



Contents

SPECIAL REPORT

Making the move from
physics to finance

p220

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

Each year, some scientists switch careers to become consultants. But can they weather the current economic storm? Tobias Kramer and Amy Wilson investigate.

On a late September day five years ago, the champagne corks were popping in Katrin Suder's lab at the Ruhr University in Bochum, Germany. That day, her group's paper on neurophysiological mechanisms of visual perception had been accepted for publication. The physicist's expertise in mathematical modelling opened the doors to a successful career in academic

science. But instead, Suder turned her skills to a career in financial analysis.

Suder now works in the German office of international management consultants McKinsey. She has found that the methods and tools of physics translate well into the business world. At university, her research involved building computer models of visual perception. Now, instead of modelling biological phenomena, she uses the same computational and mathematical skills to model factors such as consumer response.

Five years on, those who hope to follow in Suder's footsteps may find the transition a little harder to make. In the United States, prospects are shakier in the wake of the 11 September 2001 terrorist attacks, after which the stock market plunged and Wall Street shed over 25% of its employees. In Europe, a sluggish economy has resulted in higher unemployment for physicists, regardless of sector. But opportunities remain on both sides of the Atlantic for physicists who are nimble and can apply their skills to multiple markets.

Some physicists may still be able to find jobs by applying for positions at large firms such as McKinsey. Others choose to go off on their own, using their skills and contacts to forge an independent career. And, most recently, these have been joined by laid-off scientists scrambling to set themselves up as consultants.

McKinsey, which has a reputation for recruiting PhD

scientists, spends time and money teaching its new employees the language of economics. Over half of its staff don't have a financial background, so the firm offers them a crash-course in economics.

But over the past six months, the company has seen lay-offs and a recruitment freeze. Wolfgang Pleyer, managing director of the financial-consulting company d-fine in Eschborn, Germany, thinks that the current recruitment freeze is only temporary. When it ends, he believes that physicists will be well positioned to



Alternative path: for physicists, the financial sector offers career paths that can make good use of the skills they have developed doing their research.

PHOTODISC

J. SOHM/CORBIS

get jobs, as their skills in mathematics and modelling can be broadly applied to many economic and financial models. At d-fine, for example, physicists make up 70% of the workforce.

There are some bright spots in the European market. For instance, the situation among more traditional disciplines in industrial physics isn't too bleak, says Roland Sauerbrey, head of the German Physical Society. And that could mean opportunities for consultants with specialized expertise, especially in growth areas such as nanotechnology and photonics. The supply-and-demand situation in Europe is also favourable. The UK-based Institute of Physics reports an increase in production levels for physics-based industries of between 12% in Germany and 27% in France. Meanwhile, fewer students are earning physics PhDs.

SEEKING INDEPENDENCE

For many scientists the switch to business consulting is not so much a choice but a necessity as the sagging economy sees them getting laid off. For others it is more of a lifestyle choice and an opportunity to become their own boss.

In 1999, Jack Owicki chose to leave a successful career as a biophysics professor and use his wide range of contacts and broad career focus to become a consultant for pharmaceutical and instrumentation companies. He focuses on detection technologies, in particular the use of fluorescence in high-throughput screening.

After working as a tenured physics professor at the University of California, Berkeley, and then at instrumentation companies Molecular Devices and LJL Biosystems, Owicki decided he wanted to take control of his life and he struck out on his own. "It was a career choice as opposed to a holding pattern until I found a job," he says. In addition to his consultancy work, he teaches courses in optical detection systems. These courses are offered at national conferences, but are primarily given to companies wishing to provide their employees with additional knowledge.

Like others, Owicki has had to make some adjustments to his business strategy because of the changing economy. In the past, he was successful simply relying on word-of-mouth recommendations, but now he



Challenging economic conditions have led many physicists to consider a career change away from industry.

finds that he needs to advertise and has to travel more than he would like — although he is still as busy as he wants to be.

So what does it take to be a successful consultant? Al Kolb, who, like Owicki, left academia to become an independent consultant, says that you "have to have a host of contacts in order to make it work". The first step is to make an honest assessment of your expertise and what you have to offer. He also stresses the importance of developing your business networking and social skills.

Kolb got his PhD in molecular biology and biochemistry at the University of California, Los Angeles, and has spent several years in industry working in instrumentation and automation for Beckman Coulter, Packard Instrument and the Automation Partnership. He says that he has witnessed an amazing growth in the drug-discovery industry in the 1990s, but has seen a shift in priorities in recent years. Many companies have been forced to abandon early-stage screening efforts and now limit their focus to late-stage drug discovery. This will

eventually have to shift back as drug pipelines dry up, Kolb says. He feels confident that the industry will rebound, but for now, as companies have fewer and fewer resources of their own, they will turn to consultants to assist them in short-term projects or strategic development.

Financial markets are cyclical. There are already signs of a stock market recovery. An astute physicist could model the markets, measure the opportunity and start a new career when the time is right. ■

Tobias Kramer was an intern at Nature's Munich office and is now a postdoc in theoretical physics in Munich; Amy Wilson is a technical consultant for law firm Stroock & Stroock & Lavan in New York.



Physicist Katrin Suder opted out of an academic career to use her scientific talents as a consultant.